PATENT ABSTRACTS OF JAPAN

(11)Publication number:

2000-285140

(43) Date of publication of application: 13.10.2000

(51)Int.CI.

G06F 17/30

(21)Application number: 11-343890

(71)Applicant: RICOH CO LTD

(22)Date of filing:

02.12.1999

(72)Inventor: SHIMADA ATSUO

MIYAJI TATSUO **KENMOCHI EIJI**

YAMAZAKI MAKOTO TAKEYA KAZUHISA NAGATSUKA TETSUO

(30)Priority

Priority number: 10376576

Priority date : 24.12.1998

Priority country: JP

10369589

25.12.1998

29.01.1999 11022915

JP

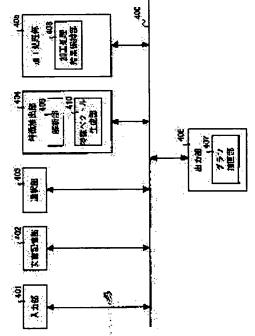
JP

(54) DEVICE AND METHOD FOR PROCESSING DOCUMENT, DEVICE AND METHOD FOR CLASSIFYING DOCUMENT, AND COMPUTER READABLE RECORDING MEDIUM RECORDED WITH PROGRAM FOR ALLOWING COMPUTER TO EXECUTE THESE METHODS

(57)Abstract:

PROBLEM TO BE SOLVED: To make it possible not only to support the result but also to support the whole information analysis in the case of analysis related to the meaning of a document.

SOLUTION: This document processing device is provided with a document storage part 402 for storing inputted document data, a selection part 403 for selecting all the parts or one part of the document data stored in the document storage part 403, a feature extracting part 404 for extracting data on the features of character strings from all the parts or one part of the document data selected by the selection part 403, a work processing part 405 for working all the parts or one part of the document data on the bases of the data on the features of character strings extracted by the feature extracting part 404 and an output part for outputting all the parts or one part of the document data worked by the work processing part 405.



LEGAL STATUS

[Date of request for examination]

8 炒 噩 揷 唧 Þ 垬

(19)日本国特許庁(J P)

3 (11)特許出購公開番号 特開2000-285140

(43)公開日 平成12年10月13日(2000.10.13) (P2000-285140A)

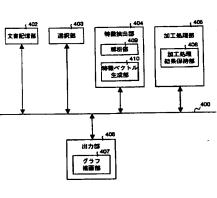
(51) Int.Q. G06F 17/30 **数**別記号 G06F FI 15/401 15/401 15/40 370A 3 1 0 D テーマコート・(参考) 5B075

警任網头 未請求 請求項の数38 OL (全の頁)

是共同に扱く		1	(SS) DEJUMENT DE LA
		1 * (IP)	(33) 阿托索士提回
	29)	平成11年1月29日(1999.1.29)	(32)優先日
弁理士 酒井 昭高		特職平11-22915	(31)優先権主張番号
	(74)代理	日本 (JP)	(33)優先權主張国
会社リコー内	25)	平成10年12月25日(1998, 12, 25)	(32)優先日
東京都大田区中周込1丁目3番6号 探式		特闘平10-369589	(31)優先権主張番号
宮樹 選生	(72)発明者	日本 (JP)	(33)優先権主張国
会社リコー内	24)	平成10年12月24日(1998, 12, 24)	(32)優先日
東京都大田区中馬込1丁目3番6号 株式		特顯平10-376576	(31)優先権主張番号
場田 教夫	(72)発明者		
東京都大田区中馬込1丁目3番6号	2)	平成11年12月2日(1999,12.2)	(22)出版日
株式会社リコー			
(71)出額人 000006747	(71)田健し	特闘平11-343890	(21)出願番号
	_		

Ē [発明の名称] 文書処理装置、文書分類装置、文書処理方法、文書分割方法およびそれらの方法をコンピュータ に契行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

05により加工処理された文書データの全部または一部 特徴に関するデータに基づいて文書データの全部または 部404と、特徴抽出404により抽出された文字列の 部から文字列の特徴に関するデータを抽出する特徴抽出 択部403により選択された文書データの全部または一 データの全部または一部を選択する選択部403と、選 徳部402と、文 記憶部402により記憶された文書 作業全般にわたる支援をおこなうことを課題とする。 て、単にその結果のみを出力するのではなく、情報分析 を出力する出力部406とを備える。 一部を加工処理する加工処理部405と、加工処理部4 【解決手段】 入力された文書データを記憶する文書記 文 の意味に係わるような分析作業におい



【特許請求の範囲】

式で表示または印刷するために出力する文書処理装置に 【請求項1】 入力された複数の文書データを所定の形

前記文書記憶手段により記憶された文書データの全部ま 入力された文書データを記憶する文書記憶手段と、

前記選択手段により選択された文書データの全部または 一部から文字列の特徴に関するデータを抽出する特徴抽

前記特徴抽出手段により抽出された文字列の特徴に関す

前記加工処理手段により加工処理された文書データの全

を備えたことを特徴とする文書処理装置。

部または一部の内容に基づいて複数の項目値を設定する

文書データの全部または一部を集計する集計手段と、 **前記項目値設定手段により設定された項目値ごとに前記**

前記文書データの全部または一部を、項目値を少なくと とする請求項1に記載の文書処理装置。 も一つの軸とする表形式に展開して出力することを特徴

一夕の全部または一部とともに出力することを特徴とす を、前記加工処理手段により加工処理される前の文書で 手段により加工処理された文書データの全部または一部 る請求項1または2に記載の文書処理装置。

処理手段により加工処理された文書データの全部または 一部を記憶することを特徴とする請求項1~3のいずれ

ることを特徴とする請求項1~4のいずれか一つに記載 により出力された文書データの全部または一部を選択す

処理の内容に関するデータを記憶することを特徴とする 請求項1~5のいずれか一つに記載の文書処理装置。 【請求項6】 前記文書記憶手段は、さらに、前記加工 【請求項7】 文書の内容に基づいて文書の分類をおこ

文書データを入力する入力手段と、 前記入力手段により入力された文書データを解析して言

前記書語解析手段により得られた書語解析情報に基づい て前記文書データに対する文書特徴ベクトルを生成する

(2)

特期2000-285140

たは一部を選択する選択手段と、

るデータに基づいて前記文書データの全部または一部を 加工処理する加工処理手段と、

部または一部を出力する出力手段と、

前記加工処理手段により加工処理された文書データの全 項目値数定手段と、 【請求項2】 前記出力手段は、

【請求項3】 前記出力手段は、さらに、前記加工処理

か一つに記載の文書処理装置。 【請求項4】 前記文書記憶手段は、さらに、前記加工

の文書処理装置。 【請求項5】 前記選択手段は、さらに、前記出力手段

なう文書分類装置において、

語解析情報を得る言語解析手段と、

入力部

ル間の類似度に基づいて文書を分類し、文書の部分集合 前記ペクトル生成手段により生成された文 特徴ベクト

を生成する分類手段と、 前記分類手段により生成された文書の部分集合の特徴で

微を分類体系の構成要素として記憶する分類体系記憶手 前記クラスタ特徴算出手段により算出されたクラスタ特 あるクラスタ特徴を算出するクラスタ特徴算出手段と、

を備えたことを特徴とする文書分類装置。

なう文書分類装置において、 【請求項8】 文書の内容に基づいて文 の分類をおこ

文書データを入力する入力手段と

語解析情報を得る言語解析手段と、 前記入力手段により入力された文 データを解析して言

前記言語解析手段により得られた 語解析情報に基づい て前記文書データに対する文 特徴ベクトルを生成する ベクトル生成手段と、

前記ペクトル生成手段により生成された文書特徴ベクト ル間の類似度に基づいて文 を分類し、文 の部分集合 を生成する分類手段と、

前記分類手段により生成された文 の部分集合の特徴で 前記分類手段により生成された文 の部分集合の中から 徴を表示する表示手段と、 あるクラスタ特徴を算出するクラスタ特徴算出手段と、 所望の部分集合を選択するクラスタ選択指示手段と、 前記クラスタ特徴算出手段により算出されたクラスタ特

集合を分類体系の構成要素として記憶する分類体系記憶 前記クラスタ選択指示手段により選択された文 の部分

を備えたことを特徴とする文 分類装置。

文書特徴ベクトルを記憶する文 特徴ベクトル記憶手段 【請求項8】 前記ペクトル生成手段により生成された

たのこりとなるように 正するベクトル停正手段と れた部分集合に属する文 の文 特徴ペクトルを除去し 微ペクトルを、前記クラスタ選択指示手段により選択さ 前記文書特徴ペクトル記憶手段により記憶された文 特

た文書特徴ペクトルに基づいて文書を分類することを特 前記分類手段は、前記ベクトル修正手段により修正され

た文書特徴ベクトルを記憶する文 特徴ベクトル記憶手 微とする請求項8に記載の文 分類装置。 【請求項10】 前記ペクトル生成手段により生成され

徴ベクトル間の類似度を判断する際の文 表現空間を前 前記文書特徴ベクトル記憶手段により記憶された文 特 算出する特徴量に基づいて修正する文 表現空間修正手 記クラスタ選択指示手段により選択された部分集合から

前記分類手段は、前記文 表現空間修正手段により修正された文書表現空間を用いて、前記ペクトル生成手段により生成された文 特徴ペクトル間の類似度に基づいて文 を分類することを特徴とする請求項8に記載の文書分類装置。

(請求項11) 前記ペクトル生成手段により生成された文 特徴ペクトルを記憶する文書特徴ペクトル記憶手段と、

前記文(特徴ペクトル記憶手段により記憶された文書特徴ペクトル間の類以度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文 表現空間修正手段により修正された文 表現空間を用いて、前記ペクトル修正手段により修正された文 特徴ペクトル間の類似度に基づいて文 を分類することを特徴とする請求項9に記載の文書分類装置。

【請求項12】 前記分類手段により生成された文書の 部分集合に所属する文 のすべてあるいは一部が選択さ れた場合に選択されたことを示す選択情報を付与する選 択情報付与手段を備え、

前記表示手段は、前記グラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする請求項8または10に記載の文 分類装置。

【請求項13】 約記分類体系記憶手段は、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文 のほか、クラスタ特徴および/または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする請求項8~12に記載の文公報技術

【請求項14】 文 の内容にしたがって文書群を分類する文 分類装置において、

文 データ群を入力する文書入力手段と、

入力された文 データ群の各文書に対して所定の基準に 基づき文 の分割をおこない、一つの文書データからー つまたは複数の分割文 データを生成する文書分割手段 と、

前記文 データと前記分割文書データとの対応を示す文 一分割文 対応マップを生成する文書一分割文書対応 マップ年成手段と、

前記分割文 データを分類する分割文書分類手段と、 前配分割文 分類手段による分類結果に基づいて分割文 分類結果情報を生成する分割文書分類結果生成手段

前記文 --分割文 対応マップと前記分割文書分類結果 情報とを用いて前記文 データの分類結果情報を生成す る文 分類結果生成手段と、

5

を備えたことを特徴とする文書分類装置。 『辞求頃15】 前記文書データを保存する』

【線求項15】 前記文書データを保存する文書保存年 のに

前記分割文書データを保存する分割文書保存手段と、 前記文書 一分割文書対応マップ生成手段により生成され た文書 一分割文書対応マップを保存する文書 一分割文書 対応マップ保存手段と、

を備えたことを特徴とする請求項14に記載の文書分類 結響

【請求項16】 前記分割文書分類結果生成手段により 生成された分割文書分類結果情報を保存する分割文書分 類結果保存手段を備えたことを特徴とする請求項15に 記載の文書分類装置。

(請求項17) 前記文書分割手段により生成される複数の分割文書データには分割前の文書データをひものを 数の分割文書データには分割前の文書データそのものを 含むことを特徴とする請求項14~16のいずれか一つ に記載の文書分類装置。

【請求項18】 前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする請求項14~17のいずれか──つに記載の文書分類装置。

【請求項19】 前記文書データに含まれる要素を抽出 する文書要素抽出手段と、

前記文書要素抽出手段により抽出された要素に付随する 要素付随情報を抽出する要素付随情報抽出手段と、

前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付随情報抽出手段により抽出された要素付随情報とを用いて前記文書データを分割する構成にしたことを特徴とする講求項14~17のいずれか一つに記載の文書分類装置。

【請求填20】 前記文書分割手段が、指示された指定 範囲にしたがって文書データの分割をおこなう構成にし たことを特徴とする請求項14~17のいずれか一つに 記載の文書分類装置。

【請求項21】 前記文書分割手段が、文書データ中の 文字数、文数、または文字数と文数の両方を基に文書デ ータを分割する構成にしたことを特徴とする請求項14 ~17のいずれかーつに記載の文書分類装置。

【請求項22】 前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする請求項14~21のいずれか一つに記載の文書分類装置。

【請求項23】 前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする請求項22に記載の文書

【請求項24】 入力された複数の文 データを所定の

形式で表示または印刷するために出力する文書処理方法 ・ において、

入力された文書データを記憶する文書記憶工程と、 前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、

前記選択工程により選択された文書データの全部または 一部から文字列の特徴に関するデータを抽出する特徴抽 出工程と、

前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、

前記加工処理工程により加工処理された文書データの全 師または一師を出力する出力工程と、 た金ノゼートを独飾とするできの細方は

を含んだことを特徴とする文書処理方法。

【請求項25】 前記出力工程は、 前記加工処理工程により加工処理された文書データの全 虧または一部の内容に基づいて複数の項目値を設定する 確日は400~550~

前記項目値設定工程により設定された項目値ごとに前記 文書データの全部または一部を集計する集計工程と、 た今ユ

前記文書データの全部または一部を、項目値を少なくとも一つの動とする表形式に展開して出力することを特徴とする表形式に展開して出力することを特徴とする請求項24に記載の文書処理方法。

【請求項26】 前記出力工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする請求項24または25に記載の文章処理方法。

【請求項27】 前記文書記憶工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする請求項24~26のいずれか一つに記載の文書処理方法。

(請求項28) 約記選択工程は、さらに、前記出力工程により出力された文書データの全部または一部を選択することを特徴とする請求項24~27のいずれか一つに記載の文書処理方法。

【請求項29】 前記文書記憶工程は、さらに、前記加工処理の内容に関するデータを記憶することを特徴とす る請求項24~28のいずれか一つに記載の文書処理方は、

【請求項30】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、 前記入力工程により入力された文書データを解析して言 語解析情報を得る言語解析工程と、

前配言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトルを生成するベクトルを生成するベクトル生成工程と、

前配ベクトル生成工程により生成された文 特徴ベクト

ル間の類似度に基づいて文 を分類し、文書の部分集合 を生成する分類工程と、

前記分類工程により生成された文 の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、 が記クラスタ特徴算出工程により算出されたクラスタ特 前記クラスタ特徴算出工程により算出されたクラスタ特 徴に基づいて分類体系の構成要素を生成する分類体系生

を含んだことを特徴とする文 分類方法。

【餅求項31】 文書の内 に基づいて文 の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、 前配入力工程により入力された文書データを解析して言

語解析情報を得る實語解析工程と、 前記書語解析工程により得られた書語解析情報に基づい て前記文書データに対する文 特徴ベクトルを生成する

ベクトル生成工程と、 前記ベクトル生成工程により生成された文書特徴ベクト ル間の類似度に基づいて文 を分類し、文書の部分集合を生成する分類工程と、

在主席9のの地土性に、 前記分類工程により生成された文 の部分集合の特徴で あるクラスタ特徴を算出するクラスタ特徴算出工程と、 前記クラスタ特徴算出工程により算出されたクラスタ特 機を表示する表示工程と、

前配分類工程により生成された文の部分集合の中から所望の部分集合を選択するクラスタ選択指示工程と、所望の部分集合を選択するクラスタ選択指示工程と、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文 分類方法。

(請求項32) 前記ペクトル生成工程により生成された文書特徴ペクトルを、前記クラスタ選択指示工程により選択された部分集合に属する文 の文 特徴ペクトルを除去したのこりとなるように修正するペクトル修正工程と、

を含み、 前記分類工程は、前記ベクトル修正工程により修正され た文書特徴ベクトルに基づいて文 を分類することを特

徴とする請求項31に記載の文 分類方法。 【請求項33】 前記ペクトル生成工程により生成され た文書特徴ペクトル間の類似度を判断する際の文 表現 空間を前記クラスタ選択指示工程により選択された部分 集合から算出する特徴量に基づいて修正する文 表現空 間修正工程と、

前記分類工程は、前記文 表現空間棒正工程により棒正された文書表現空間を用いて、前記ペクトル生成手段工程により生成された文 特徴ペクトル間の類似度に基づいて文書を分類することを特徴とする請求項31に記載の文書や類方法。

【請求項34】 前記ペクトル生成工程により生成され

6)

た文 特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間終正すおす

を含み、

前記分類工程は、前記文 表現空間修正工程により修正 された文 表現空間を用いて、前記ペクトル修正工程に より修正された文 特徴ペクトル間の類似度に基づいて 文 を分類することを特徴とする請求項32に記載の文

(請求填35) 前記分類工程により生成された文書の部分集合に所属する文 のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、

前記表示工程は、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする請求項31または33に記載の文 分類方法。

【請求項36】 前記分類体系生成工程は、前記選択指示工程により選択されたクラス与特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文 昇の全部あるいは一部および/または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする請求項31~35に記載の文書分類方は

【請求項37】 文 の内容にしたがって文書群を分類する文 分類方法において、

文書データ群を入力し、入力された文書データ群の各文 に対して所定の基準に基づき文書の分割をおこない、この文 データから一つまたは複数の分割文書データ を生成し、前記文 データと前記分割文書データとの対 応を示す文 一分割文書対応マップを生成し、前記分割 文 データを分類し、分割文書分類結果情報を生成し、前記文 一分割文 対応マップと前記分割文書分類結果情報を生成し、前記文 一分割文 対応マップと前記分割文書分類結果情報を生成す情報とき用いて前記文書データの分類結果情報を生成することを特徴とする文 分類方法。

【錦水項38】 前記請求項24~37のいずれか~つ に記載された方法をコンピュータに実行させるプログラ 人を記録したことを特徴とするコンピュータ読み取り可 館な記録媒体。

【発明の詳細な説明】

[0001]

【発明の属する技術分野】この発明は、入力された複数の文 データを所定の形式で表示または印刷するために出力する文 処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。また、この発ュータ読み取り可能な記録媒体に関する。また、この発明は、入力された複数の文書をその文書の内容に基づいて分類をおこなう、特に文 分類の際に算出される分類カテゴリ(体系)を精錬化する文 分類装置、文 分類

方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【従来の技術】近年、さまざまな文書分類装置や文書検 索装置が開発されている。また、インターネット等のネ ットワーク技術の普及により国内外の大量の電子化文書 ・ハワーク技術の普及により国内外の大量の電子化文書 ・ハワクセスが可能になり、それに比例して業務上電子 的に蓄積される情報の量も飛躍的に拡大した。その中で 収集した大量の文書情報を意味あるカテゴリ(体系)に 分類する等の知的作業の必要性が高まってきている。

【0003】これらの大量の文書情報を意味的に分類するという作業の目的は、以下のようなものである。まず第1に、検索容易性の向上が考えられる。これは、膨大な文書群を分類名称(内容名)を手がかりに検索できるので検索が比較的容易になるというものである。

【0004】第2に、情報群全体の把握が考えられる。これは、文書群全体がどのような内容(個々の分類)で 構成されているかを把握する。しかし、大量の文書情報 構成されているかを把握する。しかし、大量の文書情報 を操作者が手動で分類する場合、正確な分類をすること はできるが、分類に張る人的・時間的コストが膨大なも にできるが、分類に張る人的・時間的コストが膨大なも のになるため、近年の文書の蓄積量の膨大さから、文書 情報の自助分類装置が提案されるようになってきた。

【0005】文書自助分類装置の従来技術としては、たとえば、特開平フー36897号公報に記載されているように、文書を、単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルにま分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【0006】 対た、「Projections for Efficient Document Clustering (革命名: Hinrich Schutzeend Craing Silverstein.

いる。そのため、大量の文書群を意味のあるカテゴリに り、大量の文書群へのアクセスが可能になり、その結 は、確率論的アプローチを用いる方法等が考えられる。 分類を実施しているものがある。そのほかの方法として 1997)」においては、潜在的意味空間において文書 of SIGIR. ページ:74-81. 発行年: 学会名:ACM. 論文名:Proceedings によりおこなう場合、その人的および時間的なコストが なわれ始めている。しかし、このような分類作業を人手 分類し、文書群の構造を把握するという知的作業がおこ かつ、効率的に利用できるようにする必要性が高まって 果、その文書群をさまざまな利用者の意図に基づいて、 類基準も変わってしまうことになる。 のみが有することになるため、分類担当者が代わると分 膨大なものになるし、また、分類のための知識を分類者 【0007】また近年、インターネットなどの普及によ

【0008】そのため、文 群を人間が分類するような

分類基準で自動的に分類しうる文書分類装置が望まれており、文書分類装置としては、たとえば、特界平フー114572号公報に記載されているように、文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルをもとに文書分類することで、意味的な異なりを用いたももとに文書分類することで、意味的な異なりを用いた自動分類を可能にするものがある。

000

【発明が解決しようとする課題】しかしながら、上記従来技術の文書分類装置は、本質的には単語で構成される多次元空間に布置した文書を統計的な分類をする方法であるため、分類結果は単語のいわゆる振る舞いという観点から統計的に求められたものにすぎず、分類の結果、算出される各クラスタ(分類された個々の文書の部分集合)が操作者(利用者)に理解不能な場合がある。

【0010】また、どのような分類結果が最適かは、分類対象の文書集合の特徴や、利用者の作業の目的に依存するため、最適な分類結果について定義することが困難であるという問題点があった。特に、上記情報群全体の把値に関し、多様な操作者の意図により要求される分類も異なるため、一度の分類作業で、操作者の所望する結果を得ることが困難であるという問題点があった。

【0011】このように、文書分類の結果は、多くのいわゆるノイズを含んだものであると解釈することができ、その一部についてのみが操作者にとって有益な場合が多いという問題点があった。

[0012]また、これらの従来技術においては、文書の構成単位を考慮していないため、文書が一つまたは複数の段落記号やタイトルなどにより区切られた構造を持つ場合には、一つの文書の中に複数の話題や意味が含まれてしまい、その結果、利用者がその分類カテゴリを理解し難くなったり、また、ある特定の話題や特定の意味に限定されたカテゴリになったり、利用者の意図するカテゴリとは異なるカテゴリに分類されてしまうという問題か生じている。

【0013】なお、特別平6-176064号公報に示された文脈依存自助分類装置には、文書の段落情報を考慮した文書自助分類をおこなうことにより分類精度を高めた文書自助分類をおこなうことにより分類精度を高めようとするものが関示されているか、本質的に上記の問題を解決するものではない。

【0014】また、上記従来技術の文書分類装置や文書検索装置等の文書処理装置は、単に文書を分類する、あるいは文書を検索する機能を有するのみで、その結果を用いてさらなる分析をおこない、文書群に内在する隠れた情報の解析をおこなうことについては何ら考慮がされておらず、文書群に内在する隠れた情報の解析は別の解析は別の解析をおこなわなければならないという問題点が装置を用いておこなわなければならないという問題点があった。

【0015】また、情報分析をおこなう操作者が分類作 乗や検索作業をおこなうのは、これらの作業において、 結果は目的なのではなく、単に情報分析作業の途中経過

にすぎないからである。通常は、その後、さらに結果を把握しやすくするために、元の文書に含まれる情報を最大限に活用し、結果の並べ替えをおこなったり、集計・大限に活用し、結果の並べ替えをおこなったり、集計・収置を施したり、結果をもとに表の形式にまとめたり、さらにはグラフ化したりというようなさまざまな処り、さらにはグラフ化したりというようなさまざまな処理を繰り返しおこない、意味ある情報分析結果を導き出

【0016】また、数値データを対象とする情報の分析作業において、表計算ソフトウェアが用いられる場合があるが、表計算ソフトウェアは、元来、数値データの取扱いを裁回して開発されたものであり、文字データ、特に文書の意味に張わるような分析作業においては十分な効果を発揮することはできなかった。

【0019】またこの発明は、上述した従来例による問題点を解決するため、一つの文 の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されることがないことにより、利用者かその分類カテゴリをよく理解できる文分組装置、文書分割方法およびその方法をコンピュータに実行させるプログラムを記録にたコンピュータ語の別可能な記録媒体を提供することを第3の目的とする。

【課題を解決するための手段】上述した課題を解決し、目的を達成するため、請求項1の免明に係る文 処理装置は、入力された複数の文 データを所定の形式で表示または印刷するために出力する文 処理装置において、入力された文書データを記憶する文 記憶手段と、前記及状手段により記憶された文 データの全部または一部を選択する選択手段と、前記選択手段により超扱された文書データの全部または一部を選択する選択手段と、前記選択手段により選択された文書データを抽出する特徴出手段と、前記特徴抽出手 支 ボデータを抽出する特徴出門するデータにあった文字列の特徴に関するデータを抽出する特徴出手段により抽出された文字列の特徴に関するデータにあった。

8

備えたことを特徴とする。

【0021】この請求項1の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0022】また、請求項2の発明に係る文書処理装置は、請求項1の発明において、前記出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記項目値設定手段により設定された項目値ごとに前記文 データの全部または一部を集計する集計手段と、を備え、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする。

【0023】この請求項2の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0024】また、請求項3の免明に係る文書処理装置は、請求項1または2の免明において、前記出力手段が、さらに、前記加工処理手段により加工処理された文 データの全部または一部を、前記加工処理手段により加工処理される前の文 データの全部または一部とともに出力することを特徴とする。

【0025】この請求項3の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味にほわるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0026】また、請求項4の発明に係る文書処理装置は、請求項1~3の発明において、前記文書記憶手段が、さらに、前記加工処理手段により加工処理された文データの全部または一部を記憶することを特徴とする

【0027】この請求項4の発明によれば、以後、他のデータと同様に扱うことができることから、文書の意味に保わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0028】また、請求項5の発明に係る文書処理装置は、請求項1~4の発明において、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択することを特徴とする。

らに、前配出力手段により出力された又書データの全部または一部を選択することを特徴とする。
【0029】この請求項5の発明によれば、出力手段により出力された文 データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作

東ができることから、文書の意味に係わるような分析作 集において、単にその結果のみを出力するのではなく、 情報分析作業全般にわたる支援をおこなうことができ

【0030】また、請求項6の発明に係る文書処理装置は、請求項1~5の発明において、前記文書記憶手段が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【0031】この請求項6の免明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0032】また、請求項7の免明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析機を得る言語解析手段により前記べクトルを生成するベクトルを生成するベクトルを生成するベクトルを生成するベクトルを生成された文書特徴ベクトルの記入を手段により生成された文書特徴ベクトルの研究の度に基づいて文書の部分集合を生成する分別手段と、前記クラスタ特徴を算出するクラスタ特徴を算出するクラスタ特徴があるクラスタ特徴を算出するクラスタ特徴が表記憶手段と、前記クラスタ特徴を算出するり、対して表さる。

【0033】この請求項7の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

類装置において、文書データを入力する入力手段と、前 は、文書の内容に基づいて文書の分類をおこなう文書分 の部分集合を選択するクラスタ選択指示手段と、前記ク を生成する分類手段と、前記分類手段により生成された 前記ペクトル生成手段により生成された文書特徴ペクト 解析情報を得る言語解析手段と、前記言語解析手段によ 記入力手段により入力された文書データを解析して言語 ラスタ選択指示手段により選択された文 の部分集合を 分類手段により生成された文書の部分集合の中から所望 り算出されたクラスタ特徴を表示する表示手段と、前記 ラスタ特徴算出手段と、前記クラスタ特徴算出手段によ 文書の部分集合の特徴であるクラスタ特徴を算出するク ル間の類似度に基づいて文書を分類し、文書の部分集合 する文書特徴ベクトルを生成するベクトル生成手段と、 り得られた言語解析情報に基づいて前記文書データに対 【0034】また、請求項8の発明に係る文書分類装置

> 分類体系の構成要素として記憶する分類体系記憶手段 と、を備えたことを特徴とする。

【0035】この臍求項8の発明によれば、選択された クラスタのみを用いており、より操作者の意図したもの に近いクラスタの構造化・体系化をおこなうことができ ス

【0036】また、請求項9の発明に係る文書分類装置は、請求項8の発明において、前記ペクトル生成手段により生成された文書特徴ペクトルを記憶する文書特徴ペクトル記憶手段により記憶された文書特徴ペクトルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ペクトルを正手段と、を備え、前記分類手段が、前記ペクトル修正手段により修正された文書特徴ペクトルに基づいて文書を分類することを特徴とする。

【0037】この請求項9の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0038】また、請求項10の売明に係る文書分類装置は、請求項8の売明において、前記ペクトル生成手段により生成された文書特徴ペクトル記憶する文書特徴ペクトル記憶手段により記憶された文書特徴ペクトル間の類似度を判断する際の文書表現空間を前記グラスタ選択指示手段により選択された部分集合から算出する特徴長しまういて作品・リ表を表現空間修正手段に、非正を力に大力達表現空間修正手段により作正された文書表現空間修正手段により作正された文書表現空間修正手段により年正された文書表現空間修正手段により年正された文書表現空間修正手段により年正された文書表表で、「同志用いて、前記ペクトル間の類似度に基づいて文書を分類することを特徴とする。

【0039】この講求項10の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状を次回の分類実行時に排除することができる。

【0040】また、請求項11の発明に係る文書分類装置は、請求項9の発明において、前記ペクトル生成手段(こより生成された文書特費ペクトルを記憶する文書特費、ペクトル記憶手段と、前記文書特費ペクトル配信手段により記憶された文書特費ペクトル間の類似度を判断する際の文書表現空間を正年段と、を備え、前記分類手段が、お記文書表現空間修正手段により修正された文書表現空間修正手段により修正された文書表現空間修正手段により修正された文書表現空間修正手段により修正された文書表現空間を用いて、前記ペクトル修正等段により修正された文書表現公、前記べ書表現で間を用いて、前記ペクトル側の類収度に基づいて文書を分類することを特徴ペクトル間の類収度に基づいて文書を分類することを特徴とする。

【0041】この請求項11の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の

配億手段 分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0043】この請求項12の発明によれば、多量に利用される文書の魏別性および一度も選択されない文書の魏別性および一度も選択されない文書の魏別性を向上させることができる。

ることを特徴とする。

【0044】また、請求項13の発明に係る文 分類装置は、請求項8~12の発明において、前記分類体系記億手段が、前記選択指示手段により選択された文 の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および/または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする。

簡易に生成できるので、分類体系の利用価値を向上させることができる。 【0046】また、請求項14の発明に係る文書分類装置は、文書の内容にしたがって文 群を分類する文 分 の内容把握を容易にし、かつ、操作者独自の分類体系を

【0047】この請求項14の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文 (所別文書)中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【0048】また、請求項15の発明に係る文 分類装置は、請求項14の発明において、前配文 データを保存する文 保存手段と、前記分割文 データを保存する

分割文 保存手段と、前記文書一分割文書対応マップ生成手段により生成された文書一分割文書対応マップを保存する文 一分割文 対応マップ保存手段と、を備えたことを特徴とする。

【0049】この請求項15の発明によれば、分割文書データおよび文 一分割文書対応マップを再生成することなしに、同一の文 データに対して、分類数、分類手法、または分類時の階段定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータか保存されることにより、利用者が分類作業に対して時間的な自由度を持つことができるし、過去に行った文書分類の再分析を任意の時間間におこなうこともできる。

【0050】また、請求項16の発明に係る文書分類装置は、請求項15の発明において、前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文 分類結果保存手段を備えたことを特徴とする。

【0051】この請求項16の免明によれば、請求項15の免明の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことができるし、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうこともできる。

【0052】また、請求項17の発明に係る文書分類装置は、請求項14~16の発明において、前記文書分割 手段により生成される複数の分割文書データには分割前 の文 データそのものを含むことを特徴とする。

【0053】この請求項17の発明によれば、利用者は、分割されている文 データを分類することで得られる詳細な文 データの分類構造だけでなく、分割前の文 データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることができ

【0054】また、請求項18の発明に係る文書分類装置は、請求項14~17の発明において、前記文書分割手段が、文 データの構造情報を基に文書データを分割する構成にしたことを特徴とする。

【0055】この請求項18の発明によれば、異なった 話題の分割等を適切におこなうことができ、したがっ て、文 データの詳細な分類構造がわかる文書分類を適 切におこなうことができる。

【0056】また、請求項19の発明に係る文書分類装置は、請求項14~17の発明において、前記文書デー 置は、請求項14~17の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記 文 要素抽出手段により抽出された要素に付随する要素付随情報を抽出する要素付随情報を抽出する要素付随情報抽出手段と、を備え、

٠,

前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付随情報抽出手段により抽出された要素付随情報とを用いて前記文書データを分割する構成にしたことを特徴とする。

【0057】この請求項19の発明によれば、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【0058】また、請求項20の発明に係る文書分類装置は、請求項14~17の発明において、前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたことを特徴とする。

手段が、指示されだ相定範囲にしたかって、メ争ケータの分割をおこなう構成にしたことを特徴とする。 【0059】この請求項20の発明によれば、利用者の意図に合い、かつ文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【0060】また、請求項21の発明に係る文書分類装置は、請求項14~17において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特数とする。

【0061】この請求項21の発明によれば、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【〇〇62】また、請求項22の免明に係る文書分類装置は、請求項14~21の免明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする。

易に把握することができる。
【0064】また、請求項23の発明に係る文書分類接続は、経典権のでで表明によい、対象を表現によって、表現である文書分類接続によって、表現で表の経験を表現を表現しません。

文書データの詳細な分類構造の概要や全体的な構造を容

【0064】また、請求項23の完明に係る又審分類表 置は、請求項22の発明において、前記文書分類結果生 成手段が、分割文書データを示す情報および前記分割文 書データに付随する代表的情報を、分類結果情報とし で、抽出して提示する構成にしたことを特徴とする。

【0065】この膝求填23の発明によれば、利用舎は 文書データの詳細な分類構造の概要や全体的な構造とと もにどの分割文書が起因して当該カテゴリに分類された かというようなことも容易にわかる。

【0066】また、請求項24の発明に係る文書処理方法は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理方法において、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、前記特徴抽出工程と、前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づ

いて前記文書データの全部または一部を加工処理する加工処理工程と、前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、を含んだことを特徴とする。

【0067】この請求項24の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0068】また、請求項25の発明に係る文書処理方法は、請求項24の発明において、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開せて出力することを特徴とする。

(0069)この請求項25の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把題を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にかたる支援をおこなうことができる。

【0070】また、請求項26の発明に係る文書処理方法は、請求項24または25の発明において、前記出力工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする。

【0071】この請求項26の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【〇〇72】また、請求項27発明に係る文書処理方法は、請求項24~26の発明において、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする。

【0073】この請求項27の発明によれば、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0074】また、請求項28の免明に係る文書処理方法は、請求項24~27の免明において、前記選択工程が、さらに、前記出力工程により出力された文 データの全部または一部を選択することを特徴とする。

【0075】この請求項28の発明によれば、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことがで

【0076】また、請求項29の発明に係る文 処理方法は、請求項24~28の発明において、前記文 記憶工程が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【0077】この請求項29の発明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの替理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業を担いたる支援をおこなうことができる。

【0078】また、腓東項30の毎明に係る文 分類方法は、文書の内容に基づいて文 の分類をおこなう文分類方法において、文 データを入力する入力工程と、分類方法において、文 データを入力する入力工程と、分類方法において、文 データを入力する入力工程と、 前配の力工程により入力された文 データを開析して言語解析情報を得る言語解析情報にあるいて前記文 データに対する文書特徴ペクトル生成工程により生成された文 特徴ペクトル間の類収度に基づいて文 を分類し、文 の部分集合を生成する分類工程と、前配分与スタ特徴を買出するプラスタ特徴を開出工程と、前配クラスタ特徴を買出するプラスタ特徴が出工程と、前配クラスタ特徴を買出するクラスタ特徴が出工程と、前配クラスタ特徴を買出するクラスタ特徴が出て程と、前配クラスタ特徴を買出するクラスタ特徴が出て程と、前配クラスタ特徴を開出するのラスタ特徴が出て程と、市で、方の表別体系と成立を表別を表生成する分類体系と成立を含めたことを特徴とする。

【0079】この請求項30の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

(12)

前記分類工程により生成された文書の部分集合の中から 所望の部分集合を選択するクラスタ選択指示工程と、前 記クラスタ選択指示工程により選択されたクラスタ特徴 に基づいて分類体系の構成要素を生成する分類体系生成 工程と、を含んだことを特徴とする。

【0081】この請求項31の発明によれば、選択されたグラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0082】また、請求項32の発明に係る文書分類方法は、請求項31の発明において、前配ペクトル生成工程により生成された文 特徴ペクトルを、前配ウラスタ選択指示工程により選択された部分集合に属する文書の文 特徴ペクトルを除去したのこりとなるように様正するペクトルを正工程と、を含み、前記分類工程が、前記ペクトル修正工程と、を含み、前記分類工程が、前記ペクトル修正工程により修正された文書特徴ペクトルに基づいて文 を分類することを特徴とする。

【0083】この請求項32の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0084】また、請求項33の発明に係る文書分類方法は、請求項31の発明において、前記ペクトル生成工程により生成された文 特徴ペクトル間の類似度を判断する際の文 表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文 表現空間修正工程と、を含み、前記分類工程が、前記文 表現空間修正工程により修正された文書表が、前記文 表現空間修正工程により修正された文書表 現空間を用いて、前記ペクトル生成手段工程により生成された文 特徴ペクトル間の類似度に基づいて文書を分類することを特徴とする。

【0085】この請求項33の発明によれば、前回の分類実行の結果、操作者に退択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0086】また、請求項34の発明に係る文書分類方法は、請求項32の発明において、前記ペクトル生成工程により生成された文 特徴ペクトル間の類似度を判断する際の文 表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文 表現空間修正工程と、を含み、前記分類工程が、前記文 表現空間修正工程により修正された文書表が、前記文 表現空間修正工程により修正された文書表 現空間を用いて、前記ペクトル修正工程により修正された文書表 現空間を用いて、前記ペクトル修正工程により修正された文書表 分表されることを特徴ペクトル間の類似度に基づいて文書を分類することを特徴とする。

【0087】この講求項34の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0088】また、請求項35の発明に係る文 分類方

法は、請求項31または33の発明において、前配分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、前記表示工程が、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする。

【0089】この請求項35の発明によれば、多量に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0090】また、請求項36の発明に係る文書分類方法は、請求項31~35の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および/または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする。

【0091】この請求項36の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を 簡易に生成できるので、分類体系の利用価値を向上させ え、ことができる。

【0092】また、講求項37の発明に係る文書分類方法は、文書の内容にしたがつて文書群を分類する文書分類方法において、文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成し、前記文書データと前記分割文書データとの対応を示す文書一分割文書対応マップを生成し、前記分割文書データと対策は果情報を生成し、前記文書データを分類し、分割文書分類なップと前記分割文書分類は果情報を生成し、前記文書一分割文書対応マップと前記分割文書分類結果情報を生成し、前記文書一分割文書分類は果情報を生成することを特徴とする。

【0093】この請求項37の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその 授されたりすることがなく、したがって、利用者がその分類カテゴリをよく理様できる。また、分割前文書(所属文書)中の分割文書の位置なども示されるので、利用属文書)中の分割文書の位置なども示されるので、利用属文書)中の分割文書の位置なども示されるので、利用

【0094】また、請求項38の発明に係る記憶媒体は、請求項24~37に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械脱み取り可能となり、これによって、請求項24~37の動作をコンピュータによって実現することが可能である。

[0095]

【発明の実施の形態】以下に添付図面を参照して、この 発明に係る文 処理装置、文 処理方法およびその方法

をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を推翻に粉組する。

【0096】〔実施の形態1〕まず、この発明の実施の形態1による文書処理装置を構成する情報処理システム全体のハードウェア構成を説明する。図1は、実施の形態1による文書処理装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

[0097] 図1において、実施の形態1による文書処理装置を構成する情報処理システムは、サーバー/クライアント方式で構成されている。すなわち、サーバー 101と複数のクライアント102がネットワーク103によって接続されている。クライアント102が、分類によって接続されている。クライアント102が入り類結果等の加工処理結果の表示などをおこなう。一方、クライアント102からの指示にしたがって、サーバー101は文書(テキスト)分類等の加工処理を膨大な数値演算によりおこない、その処理の結果をクライアント102へ送る。

【0098】分類処理の場合、より具体的には、サーバー101においては、テキスト分類処理(前処理、クラスタリング処理)がおこなわれ、クライアント102においては、分類データ生成、処理実行指示、テキスト分類処理(からには、分類データ生成、処理実行指示、テキスト分類結果、表示等がおこなわれる。サーバー101における処理は、表述のように、「前処理」と「分類処理」の「分かれており、その処理はデータによっては非常にの合うが大きくなる。したがって、サーバー101は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。

【0099】また、サーバー101とクライアント102との間のデータのやりとりはファイル共有という方法を用いる。すなわち、分類処理等の加工処理に用いるファイルをサーバー101上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント102からはサーバー101の共有フォルダをネットワーク共有して利用することが可能である。

【0100】つぎに、サーバー101およびクライアント102のハードウエア構成について説明する。図2は、実施の形態1による文書処理装置を構成する情報処理システムにおけるサーバー101のハードウエア構成を示す説明図である。サーバー101は、たとえばワークステーション(WS)等が用いられる。

【0101】図2において、201はサーバー101金体を制御するCPUを、202はプートプログラム等を記憶したROMを、203はCPU201のワークエリアとして使用されるRAM203を、204は通信回線205を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインットワーク103と内部のインターフェイスを司るイン

ターフェイス(I / F)を、206はデータを記憶する ディスク装置を示している。200は上記各部を結合させるためのバスを示している。

【0102】そのほか、文書情報、國像情報、機能情報等を表示するディスプレイ208や、データを入力するためのキーボード209およびマウス210等が同様に接続されていてもよい。さらに、ディスク装置206には、クライアント102との間のデータのやりとりをするための共有フォルダ207が設けられている。

【0103】また、図3は、実施の形態1による文 処理装置を構成する情報処理システムにおけるクライアント102のハードウエア構成を示す説明図である。クライアント102は、たとえばパーソナルコンピュータ(PC)等が用いられる。

[0104] 図3において、301はシステム会体を制御するCPUを、302はブートプログラム等を配慮したROMを、303はCPU301のワークエリアとして使用されるRAMを、304はCPU301のワークエリアとしたがってHD(ハードディスク) 305に対するデータのリードノライトを制御するHDD(ハードディスク) ドライブ)を、305はHDD304の制御で、き込まれたデータを記憶するHDを、306はCPU301の制御にしたがってFD(フロッピーディスク) 307に対するデータのリードノライトを制御するFDD(フロッピーディスクドライブ)を、307はFDD306の制御で書き込まれたデータを記憶する解説自在のFDの制御で書き込まれたデータを記憶する解説自在のFDを、308はドキュメント、国像、機能情報等を表示するディスプレイをそれぞれ示している。

【0105】また、309は通信回線310を介してネットワーク103に接続され、そのネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを用るインターフェイス(1人下)を、311は文字、製庫、各種指示等の入力のための年を備えたキーボードを、312はカーソルの移動や知るの様でなウインドウの移動やサイズの変更等をおこなうマウスを、313はOCR(の下1cal Character Reader) 機能を備えた固像を光学的に読み取るスキャナを、314は分類結果をもおこなうマウスを、315は正位の内容等を印刷するプリンタを、315は正記を告給を結合するためのバスをそれぞれ示している。また、HD305にはワープロソフト等のアプリケーションソフト316が記憶されている。

【0106】つぎに、実施の形態1による文 処理装置の機能的構成について説明する。図4は、実施の形態1による文書処理装置の構成を機能的に示すブロック図である。図4において、文 処理装置は、入力的401と、文書記憶的402と、選択部403と、特徴抽出部404と、加工処理部405と、出力部406を含む構成である。

【0107】人力部401、文 記憶部402、選択部

記載された命令にしたがってCPU201または301 ディスク316等の記録媒体に記録されたプログラムに たは303、あるいはディスク装置306またはハード 406は、ROM202または302、RAM203ま 等が命令処理を実行することにより、各部の機能を実現 403、特徴抽出部404、加工処理部405、出力部

スキャナ313、OCR機能を備えたスキャナ313. の文 処理装置に組み入れた場合も文書データの入力と に、そのデータベースが記録された媒体を実施の形態1 たとえば、文 データがデータベース化されている場合 ることができるものであれば、それらのすべてを含む。 た、入力部401は、上記以外に、文書データを取得す ることができる1/F204または309等である。ま またはネットワーク103を経由して文書や文書群を得 のであり、たとえば、キーボード209または311、 [0108] 入力部401は、文書データを入力するも

ら構成されており、それらの意味のあるまとまりを一つ の文 とする。また、複数の文書の集まりを、文書群と 一つ以上の文の集まりであり、文字、文字列、数値等か 【0109】ここで、文 とは、自然言語で記述された

まれていても含まれていなくてもよい。項目値は項目の 理装置の項目名と項目値の関係を示す説明図である。た れている。項目は、項目名と、項目値から構成されてい を項目名と項目値によって表現すると、図5のようにな とえば、一つの特許公報は一つの文書であり、特許公報 実際の内容である。図5は、実施の形態1による文書処 る。項目名は項目の内容を示すラベルであり、文書に含 【0110】文(は一つあるいは複数の項目から構成さ

ち一つの記憶単位に収納される。 名あるいは項目値は、文 記憶部402のセル、すなわ 憶された文 のデータ構造を示す説明図である。各項目 施の形態1による文 処理装置の文書記憶師402に記 され、文 記憶部402により記憶される。図6は、実 いは文書群は、それぞれの文書に一意な文書1Dが付与 **【0111】入力部401によって取得された文書ある**

属性値が記憶されている。 憶されている。第2番目の記憶領域602には、セルの はつぎのセルの文 記憶部402上の位置(番地)が記 領域から構成されており、第1番目の記憶領域601に 【0112】図6においては、一つのセルは3つの記憶

いうように設定することができる。第3番目の記憶領域 が「空」、「1」が「数値」、「2」が文字列・・・と いは項目値等が格納される領域の先頭位置が記憶されて 603には、セルの実際の内容、すなわち、項目名ある 【0113】セルの属性値としては、たとえば、「0」

> 変更した結果、あらかじめ確保された領域では収まり切 なる別の領域に記憶されているので、たとえば、項目を 除は、第1番目の記憶領域601に記憶されたつぎのセ 番目に記憶された記憶領域603の先頭位置を変更する 別途大きな領域を確保してそこに項目値を記憶し、第3 れなくなった場合には、セルの構造自体には影響なく、 できる。また、セルの実際の内容は、セルの構造とは異 ルの位置を変更することにより、容易に実現することが 【0114】セルの順序の並び替えや、セルの追加・削

いは項目値などが格納される領域の先頭位置が記憶され 域702には、セルの実際の内容、すなわち項目名ある には、セル属性値が記憶されている。第2番目の記憶領 の記憶領域を使用している。第1番目の記憶領域701 を示す説明図である。図7において、一つのセルは二つ の文書記憶部402に記憶された文書の別のデータ構造 【0115】図7は、実施の形態1による文書処理装置

生した場合には、記憶内容の移動操作が必要となる。 造では、セルの順序の並び替え、セルの追加・削除が発 り合うつぎの記憶領域に記憶されている。このデータ構 【0116】つぎのセルは、文書記憶部402上でとな

は光ディスク等で構成される補助記憶装置を含んでいて える半導体メモリで構成されるが、磁気ディスクあるい 【0117】文書記憶部402は、通常高速に情報を扱

は、文書記憶部402に記憶された文書あるいは文書群 の内容をセルと順次たどって読み出し、表の形式で表示 イ、液晶ディスプレイ等から構成される。出力部406 態1においては、出力部406は、CRTディスプレ は文書群は、出力部406により表示される。実施の形 【0118】文書記憶部402に記憶された文書あるい

出し、利用者の指示により棒グラフ、円グラフ、折れ線 たは印刷されたデータに基づいてグラフを描画するグラ 群の項目値に対して利用者が設定した領域の内容を読み 7は、文書記憶部402に記憶された文書あるいは文書 フ描画部407を含んでいてもよい。グラフ描画部40 グラフ等のグラフを描画し、表示または印刷する。 【0119】また、出力部406は、妻の形式で妻示ま

刷するためのプリンタ等の印刷装置を含んでいてもよ 夕、カーソルの表示等もおこなう。また、処理結果を印 に関する表示、たとえば、操作メニューやマウスポイン 【0120】出力部406は、入力部401による操作

部404へ送る。選択部403の選択方法について、図 8~図10を用いて説明する。 域のデータを文書記憶部402から読み出し、特徴抽出 者の指示により、出力部406の表示上で選択された領 【0121】選択部403は、入力部401による操作

> 故障状況の内容を示す「内容」欄806が表示される。 報対象となった自動車の年式を示す「年式」欄805. けた営業所を示す「営業所」欄803、故障情報の対象 けた日付を示す「受付日」欄802、故障情報を受け付 書ID番号を示す「番号」欄801、故障情報を受け付 は、自動車の故障状況の内容が表示された画面表示の例 となった自動車の車種を示す「車種」欄804、故障作 を示す説明図である。図8において、画面表示には、文 処理装置の出力部406による画面表示の例、具体的に 【0123】図9において、選択領域901は、矩形で 【0122】図8~図10は、実施の形態1による文書

おいても同様に、選択領域1001は、矩形で囲まれ、 表示色が変更されている部分である。 【0124】選択部403が選択する領域としては、図

囲まれ、表示色が変更されている部分であり、図10に

項目名に属する項目値全部が選択されるようにしてもよ のみ選択可能とする。 い。なお、実施の形態1では、文字列の属性を持つ領域 また、図10に示すように項目名を選択した場合はその 9に示すように、画面上の列の一部であってもよいし、

項目値の特徴が抽出される。図11は、実施の形態1に れる抽出処理の内容について説明する。選択部403に よる文書処理装置の特徴抽出部404によりおこなわれ より選択された項目値は、特徴抽出部404によりその る抽出処理の内容の一覧を示す説明図である。 【0125】つぎに、特徴抽出部404によりおこなわ

析技術等の自然言語処理技術を用いて実現する。 らの抽出処理は、規則音声合成装置や自動翻訳装置等の 字数、単語のそれぞれの出現回数... 等がある。これ 一般的に用いられている形態素解析技術あるいは構文解 る文字列に含まれる単語、その単語の単語数、単語の文 【0126】図11において、抽出処理には、対象とす

4により抽出処理された特徴量に対して、加工処理部4 われる加工処理の内容の一覧を示す説明図である。 1による文書処理装置の加工処理部405によりおこな 05により加工処理が施される。図12は、実施の形態 れる加工処理の内容について説明する。特徴抽出部40 【0127】つぎに、加工処理部405によりおこなわ

徴量を算術する「算術処理」等がある。 特徴量のうち最小値を抽出する「最小値抽出処理」、特 理」、特徴量の代表値を抽出する「代表値抽出処理」 特徴量の内容ごとに並べ替えをおこなう「並べ替え処 特徴量のうちの最大値を抽出する「最大値抽出処理」 【0128】加工処理には、同一の特徴量ごと分類する 「分類処理」、所定の特徴量を検索する「検索処理」

われる抽出された特徴量の加工処理の内容の組み合わせ 量の抽出処理の内容と、加工処理部405によりおこな 【0129】特徴抽出部404によりおこなわれる特徴 「騒音が大きい」

> て、その設定された組み合わせを操作者に提供するよう る。また、効果の高い組み合わせをあらかじめ設定し は、おのおの操作者が選択できるようにすることができ

(1

特緊2000-285140

出力部406は、加工処理結果保持部408から内容を れた加工処理結果は、出力部406により出力される。 08に保持される。加工処理結果保持部408に保持さ 理結果は、加工処理部405内の加工処理結果保持部4 読み出し、画像表示や印刷出力をおこなう。 【0130】加工処理部405により加工処理された処

る特徴 (量) として、項目値に含まれる単語それぞれの る加工処理として、分類処理を選択した場合について訪 出現回教を選択し、加工処理部405によりおこなわれ 【0131】ここで、特徴抽出部404により抽出され

の文書を分類した場合、それぞれの分類カテゴリには意 の文書を構成する単語の出現頻度が等しい場合、それら 味の近い文書が所属すると考えることができる。 る。すなわち、ある文 での単語の出現回数は、その文 二つの文書の意味は似通っていると考えることができ る。したがって、単語の出現回数を特徴量として、複数 書の意味に関係の深い特徴量であると考えることができ 【0132】一概的に、二つの女(があり、それら二つ

は複数の項目値は、特徴抽出部404に含まれる解析部 すべての項目値に対する単語1Dごとの出現回数を計数 し、一つの項目値および選択部403により選択された のうち、名詞であるものに対して一意な単語IDを付与 には、その単語の品詞情報も付与される。出現した単語 409よって項目値ごとに形態素解析等の自然言語解析 をおこない、単語に分割される。また、それぞれの単語 【0133】選択部403により選択された一つあるい

る。たとえば、選択節403により選択された項目値 項目値の特徴(量)を示す項目値特徴ベクトルを生成す 生成部410は、計数された出現回数に基づいて個々の 【0134】特徴抽出曲404に含まれる特徴ベクトル

「強装が変色する」 騒音が大きい」

「海猫がはげる」 「オーバーヒートが起こる」

「バッテリーが上がる」

「排気が無い」

ようになる。また、図14には、単語とその単語IDご であった場合、各項目の特徴ベクトルは、図13に示す との出現回数を示す。

【0135】すなわち.

[1.1.1.0.0.0.0.0.0.0.0.0.0

(E)

. 1 . 1 . という特徴ベクトルが得られる。

【0136】この項目値の特徴ペクトルは、特徴抽出部404からの出力として加工処理部405へ送られる。加工処理部405においては、項目値の特徴ペクトルを用いて、分類処理をおこなう。分類処理は、まず、おのおののベクトル間の距離を計算することにより求める。距離の尺度には、たとえば内積を用いることができる。正確の尺度には、たとえば内積を用いることができる。配種の尺度には、たとえば内積を用いることができる。との処理をおこなう。この処理には、たとえばペクトルの集合をその距離に応じて、区間のベクトルの集合に分類するKーme an 無法を用いることができる。それぞれのベクトルの分類が完了したら、ベクトルに対応する項目値に対してどの分類に属するかの番号、すなわち、クラスタ番号と、項目値に対応する文書10を付与し、加工処理部405の出力とし、出力部406により出力表示をおこなう。

【0138】図15は、クラスタ番号1501を表示した画面の表示例を示す。クラスタ番号が同一番号である文 、たとえば、番号「1」および「6」の文書のクラスタ番号が「5」であり、両者が同一の分類に属することを示している。

【0139】つぎに、請求項2の発明においてクロス表を出力する場合について説明する。入力部401により、分析対等とする文 群を読み込んだ後、操作者は分類処理をおこなう対象となる項目名、クロス表の損触あるいは緩動となる項目名、いくつに分類をおこなうかの分類数を指示する。

【0140】図16はケロス要作成のための指示回画である。図16において、指示回画1600は、処理対象項目名入力欄1601と、軸となる項目名入力欄1602と、機動指定ボタン1603と、機動指定ボタン1604と、分類数入力欄1605とから構成される。

【0141】処理対象項目名入力欄1601には、処理対象となる項目名を入力する。キーボード209等から入力するあるいは処理対象となる項目段補を表示させその中からマウス210等により選択することにより項目名を入力することができる。また、軸となる項目名入力欄1602には、軸となる項目名を入力する。入力の方法と同様である。

【0142】報報指定ボタン1603および積齢指定ボタン1604は、軸となる項目を模糊に要示させるか横軸に表示させるかを指定するためのボタンである。また、分類数入力欄1605には、いくつに分類するかその分類数を入力する。入力の方法としては、キーボード209等から数字を入力するあるいは分類数候補を表示させその中からマウス210等により選択することにより分類数を入力するようにしてもよい。

[0143] 図16においては、処理対象項目名入力欄1601には「内容」が、軸となる項目名入力欄1602には「車種」が、また、横軸指定ボタン1604がチェックされ、分類数入力欄1605には「50」が入力され、これにより、文書群の中の「内容」に基づいて、「50(圏)」に分類され、クロス表の横軸に「車種」を表示するという指示がなされていることがわかる。
[0144]クロス表作成の指示がおこなわれることにより、分類処理が実行され、その結果がクロス表で表示される。図17および図18は、分類処理の結果が表示されたクロス表を示す図である。図17において、クロス表1700は、縦軸に分類を示す「クラスタ1」、「クラスタ2」、「が表示され、横軸に車種を示す

【0145】表の縦軸、すなわち各行は、分類処理により生成されたクラスタに対応する。各行の第1欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っている。表の横軸、ずなわち各欄には、文書群の項目「車糧」に含まれる文字列が重複することなく表示される。行「クラスタ1」の各セルには、クラスタ1に分類された文書のうち、項目「車程」の値がその欄の車種と一致するものの数が表示される。

【0146】ここで、数を表示する代わりに、セルの色の濃淡や、セルを塗りつぶす面積により数の大きさを表現するようにしてもよい。また、装の最右標および最下橋には、該当する行、欄の合針が表示される。

【0147】図18において、クロス表1700のあるセルにマウスポインタ1800を移動させ、マウス210のマウスポインタ1800を移動させ、マウス210のマウスポタンを押下する、あるいはキーボード209のカーソルキー操作によりカーソルを移動させ、特定キーを押下すると、そのセルの近傍に内容表示画面18

〇1が表示されることにより、該当する文書の項目「内 容」が表示される。

【0148】内容表示回面1801には、セル内のデータ数、表示項目、セル情報、および、各データにおける表示項目の内容が表示される。マウスポインタ1800により指定されたセルにおいては、データ数:「4」、表示項目:「内容」、セル情報:「ABC2000ークラスタ1」、表示項目の内容として「内容」の4つの内容である「排気が黒い、排気が黒い、...」が表示される。これにより、マウスポインタを所望のセルに移動させてマウスポタンを押下するという簡易な操作により、セルの内容を認識することができる。

【0149】また、内容表示画面1801に表示される項目は、設定操作により変更することが可能であり、すべての項目を表示させることもでき、また、項目を選択して表示させることもできる。

[0150]各行の第1欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っているが、操作者により、この欄の書き換えをすることができる。たとえば、上記の操作によってセルの内容を確認した後、「クラスタ1」を「排気の問題」と書き換えること後、「クラスタ1」を「排気の問題」と書き換えることができる。これにより、情報内容の把題がより容易になった。

【0151】また、分類終了時に既定値としてクラスタ番号を示す文字列を入れるのではなく、そのクラスタの特徴を示す文字列を抽出し、セルに入れることも可能である。たとえば、クラスタ1に含まれる文書の項目「内容」から、もっとも頻度が高く出現する文や単語を抽出することにより実現する。

【0152】図18においては、クラスタ1には「排気が思い」あるいは「排気」等の単語が入れられる。このように、操作者は簡易な操作により文書全体の分布状態を把握するだけでなく、必要に応じて個々の文書の内容をも詳細に知ることができる。

「ABC1600」、「ABC1800」... が表示

【0153】つぎに、クロス表を作成するための出力部406の詳細な構成の内容について説明する。図19は、実施の形態1による文書処理装置の出力館406の詳細な構成を示すブロック図である。出力部406は、グラフ描画部407のほかに、項目値選定部1901、集計部1902とから構成され、集計部はさらに実際に表示する内容に対応した記憶領域を持つ表保持部1903を備えている。

【0154】項目値選定部1901は、操作者がクロス表の一つの軸として指定した項目名(軸項目名)に対して、文書記憶部402に記憶された文書データから、項目値を順次読み出し、重複のない項目値の集合を作成するに、集計部1902は、表保持部1903の項目値に対応する領域に数値を加算することにより文書の集まさせった。

【0155】つぎに、クロス表の出力手順について説明

する。図20は、実施の形態1による文・処理装置のクロス表の出力手順を示すフローチャートである。図20のフローチャートにおいて、まず、集計に先立ち、表保特部1903の内容を初期化する(ステップS200

【0156】つぎに、項目値設定的1901により作成された項目値を、表の項目値ラベルに相当する部分に当ては的(ステップS2002)、クラスタ番号を表す文字列を、クラスタ番号に相当する部分に当てはめる(ステップS2003)。

【0157】つぎに、加工処理結果保持部408に保持された項目値に対応する文 IDについて、文 記憶部402に記憶された対応する文 を参照し、その結項目402に記憶された対応する(ステップS2004)。その後、表保持部1903の対応する領域の内容に1を加算する(ステップS2005)。

【0158】すべての項目値について上記処理をおこなったか否かを判断し(ステップS2006)、すべての項目値について上記処理がおこなわれていない場合(ステップS2006否定)は、ステップS2004へ移行し、ステップS2004~S2006の処理を繰り返し

【0159】ステップS2006において、すべての項目値について上記処理がおこなわれた場合(ステップS2006肯定)は、最右列に表示するための行の合計を計算し(ステップS2007)、併せて、最下行に表示するための類の合計を計算する(ステップS200

【0160】その後、表保持部1903に構成された表を、順次読み出して出力し(ステップS2008)、すべての処理を終了する。

【0161】なお、加工処理部405から出力されたデータを、文書記憶部402に送り、文 記憶部402に他のデータとともに記憶するように構成してもよい。文書記憶部402に記憶された加工処理部405から出力されたデータは、出力部406によって表の新たな列として表示することができる。また、表の既存の列を消去し、消去した列へ上書きするようにしてもよい。

【0162】この構成では、処理の結果である加工処理 部405から出力されたデータは、文書記憶部402に おいて、今回の加工処理の対象とならなかった他のデー タと対等に取り扱うことができ、その後の分析作業等 で、もともとの入力データに存在していたか、分析作業 の途中で加工処理によって生成されたのかを区別することなく、加工処理の対象として選択することが可能であ となく、加工処理の対象として選択することが可能である。

【0163】したがって、データの性質や、おこないたい機般分析作業の内容に応じて柔軟に加工処理対象と加工処理内容を選択することができるので、多彩で高度な情報分析作業が可能となる。

(8)

【0164】また、加工処理部405への入力データとして、特徴抽出部404から出力されたデータだけではなく、選択部403により選択されたデータも含めることができる。これにより、文字列の特徴抽出を必要としないデータや、加工処理結果の数値に対してもさらなる加工処理を施すことができるので、より多彩で高度な情報分析が可能となる。

[0165] 図21~図24は、実施の形態1による文処理装置の出力部406による画面表示の別の例を示す説明図である。図21において、「番号」、「受付日」、「営業所」、「車種」、「年式」、「内容」の他に、分類処理により得られた結果である「クラスタ番号」2101が表示されている。

【0166】さらに、図21においては、選択的403により「クラスタ番号」2101が選択されており、「クラスタ番号」2101に関するデータが反転表示されている。選択された「クラスタ番号」2101をキーとして、加工処理部405により並べ替え処理をおこなことももさます。

【0167】並べ替え処理の指示により、並べ替え処理がおこなわれた結果を表示しているのが図22である。図22においては、「クラスタ番号」が「1」のものが集まって表示されるように並べ替えられ、それに続き、「クラスタ番号」が「2」のものが集まって表示されるように並び替えられる。

【0168】具体的には、「クラスタ番号」が「1」である「番号」が「2」、「11」、「15」、「23」、「35」、「54」、「63」、「73」、「82」の原で並べ替えられ、それに続き「クラスタ番号」が「2」である「番号」が「14」、「18」、「22」、「27」、「37」、...が表示されていることがわかる。

【0169】つぎに、項目「車種」の概で、「クラスタ番号」が「1」に属するものを選択する。図23においては、項目「車種」の標で、「クラスタ番号」が「1」に属するものが選択され、その選択領域2301が反転表示されていることを示している。このように、すでに「クラスタ番号」により並べ替えがおこなわれており、同一クラスタに属するものが集まって表示されているので、回面上の連続した領域として容易に選択することができる。

【0170】つぎに、選択領域2301について車糧別の発生頻度の棒グラフを表示させたのが、図24である。図24において、棒グラフ表示領域2401には、選択領域2301によって選択された「クラスタ番号」が「1」である9つの文書が選択され、その9つの文書を車糧別に棒グラフ化したものが表示される。

【0171】このように、加工処理の対象を柔軟かつ容易に選択でき、選択された対象について多様な加工処理をおこなうことができ、また、その加工処理結果も次回

の加工処理の対象とすることができるので、高度な情報 分析作象が可能となる。

【0172】このように、分類等の文字列の特徴量を抽出して、その特徴量を用いておこなう加工処理を実施した後に多種の加工処理をおこなうのを示したが、事前に多種の処理をおこなったができるようにしてもよい。「0173」に対した技術、単類」の項目を選択し、これをキーとして並べ替えをおこなった後、集まったある重種、たとえば、「4BC1600」に対して分類処理を おこなうこともできる。また、入力虧401に対して分類処理を おこなうこともできる。また、入力虧401により入力された文章が誤字等の誤りを含んでいる場合、分額等の文字列の特徴量を抽出して、その特徴量を用いて加工処理をおこなって、誤字を一括して修正し、より好適な結果をが得られるようにデータを整えることもできる。

【0174】図25は、実施の形態1による文書処理装置の文書記憶部402の詳細な構成を示すプロック図である。図25において、文書記憶部402は、設定値記憶部2501には、文書を分類する際の分類数等の分類情報記憶部2503をはじめとするさまざな設定値、すなわち文書の理装置の動作に必要な設定値に関する情報を記憶する記憶部を構えている。これにより数定値に関する情報を記憶が大量情報ともに記憶することができる。

【0175】また、設定値送受信部2502は、設定値記憶部2501によって記憶された設定値に関する情報を他の情報処理装置へ送信する。また、設定値送受信部2502は、他の情報処理装置からの設定値に関する情報を受信する。設定値送受信部2502により受信された設定値に関する情報は、設定値記憶部2501によって記憶される。

[0176]記憶された設定値に関する情報は、後に文書を再度談み込んだときに同時に談み込まれ設定値記憶的2501に記憶される。この設定値に関するに関するに対することができたり、以後の処理の際に、再利用することができる。これにより、設定値に関する情報を文書とともに保存・管理することが可能となるので、設定値に関する情報を大きとされて持切的記述なるので、設定値に関する情報の紛失を防ぎ、好適な設定値を後に再利用することができる。

【0177】図26~図28は、実施の形態1による文書処理装置の出力的406による画面表示の別の例を示す説明図である。図26において、まず、操作者が分類をおこなうへき対象である「内容」を表示画面上で選択する。それにより選択領域2601が反転表示される。つぎに、メニュー・パー2603から、分類処理ボタンつぎに、メニュー・パー2603から、分類処理ボタン2603を選択すると、分類処理に必要な分類数、すなわち、対象をいくつに分類するかについての問い合わせ画面2604が表示される。

【0178】操作者が問い合わせ回面2604において分類数を入力すると、この分類数に関する情報が文書記憶部402に記憶される。図26においては、分類数として「50」が入力されたことを示している。【0179】その後、操作者が情報分析作業を完了して、メニュー・バー2603のファイルボタン2605の選択によりボップアップする図示を省略する保存ボタンを押下すると、文書記憶部402により、操作者が指

ともに配信される。
【0180】図27において、分類結果を表示する欄2【0180】図27において、分類結果を表示する欄2701にマウスポインタ2702を移動させ、マウスポケッンを押下すると、その分類をおこなうことに用いた分類に関する情報を表示する分類情報表示画面2703が表示される。これによる分類情報表示画面2703が表示される。これにより、用いた設定値の関連づけが容易に把握することがでり、用いた設定値の関連づけが容易に把握することがでり、用いた設定値の関連づけが容易に把握することがでり、用いた設定値の関連づけが容易に把握することがでり、用いた設定値の関連が

示したファイル名が付与され、文書の情報、分類結果と

【0181】分類情報表示画面2703には、たとえば、分類に関する情報として分類がおこなわれた日時に関する情報を示す「分類日時」、分類の対象となった文書数に関する情報を示す「分類対象数」等が表示され、また、分類設定値に関する情報として、いくつに分類したかを示す「分類数」、どの品詞に基づいて分類をしたかを示す「分類品詞」等が表示される。

【0182】分類処理を実行するたびに新規な表が作成される。図28は、分類結果1を得た後、再度分類処理がおこなわれ、分類結果2が表示された状態を示している。分類結果1を再度表示させたい場合は、画面を下約のラベル上の選択領域2801ペマウスポインタを移動させ、マウスボタンを押下する。これにより、分類結果させ、マウスボタンを押下する。これにより、分類結果させ、マウスボタンを押下する。これにより、分類結果させ、マウスボタンを押下する。これにより、分類結果させ、マウスボタンを押下する。これにより、分類結果させる場合も同様の操作によりおこなうことができる。

【0183】また、図28において、各分類処理の実行に用いた設定値に関する情報が対応する表の所定の表示領域2802に表示される。この表示領域2802は、分類結果の表示を隠さないように表示させることができ、また、その表示位置を移動することもできる。これにより、分類結果と、それに用いた設定値の関連づけが容易に把鍵できる。

【0184】つぎに、実施の形態1における文書処理袋置の文書処理の一連の手順について説明する。図29は、実施の形態1による文書処理数置の文書処理の一連の手順を示すフローチャートである。

【0185】図29のフローチャートにおいて、まず、文書データが文書処理装置に入力されたか否かを判断する(ステップS2901)。ここで、文書データが入力されるのを持って、文書データが入力された場合(ステップS2901肯定)は、入力された文書データを記憶する(ステップS2902の各ステップは、文の入力がある

ごとに他のステップとは独自におこなわれるようにしてもよい。

【0186】つぎに、記憶された文 データの金部または一部が選択されたか否かを判断する(ステップS2903)。ここで、文書データの金部または一部が選択されるのを待って、選択された場合(ステップS2903特定)は、選択された文 データの金部または一部の文字列の特徴に関するデータの抽出をおこなう(ステップS2の特徴に関するデータの抽出をおこなう(ステップS200特徴に関するデータの抽出をおこなう(ステップS200特徴に関するデータの抽出をおこなう(ステップS200特徴に関するデータの抽出をおこなう(ステップS200特徴に関するデータの抽出をおこなう(ステップS200特徴に関するデータの抽出をおこなり

【0187】その後、ステップS2904において、抽出された文字別の特徴に関するデータに基づいて、分類処理等、所定の加工処理をおこなう(ステップS2905において加工処理があこなわれたデータを、表形式に展開する等の出力処理をおこなう(ステップS2906)。

【0188】さらに、ステップS2905において加工処理されてデータを元の文書データに関連づけして記憶する(ステップS2907)。また、加工処理の設定値等の加工処理の内容に関するデータも併せて記憶する(ステップS2908)。

【0189】その後、ステップS2905において加工処理されたデータの全部または一部が選択されたか否かを判断し(ステップS2908元)、選択されなかった場合(ステップS2908元)は、ステップS2900人を行し、以後、ステップS2904~S2909の処へ移行し、以後、ステップS2904~S2909の処理を繰り返しおこなう。一方、ステップS2909において、加工処理されたデータの全部または一部が選択された場合(ステップS2909肯定)は、すべての処理を終いする。

【0190】なお、実施の形態1で説明した文 処理方法は、あらかじめ用意されたプログラムをバーソナルコソピュータやワークステーション等のコンピュータで実ソピュータやアークステーション等のコンピュータで実ソディスク、フロッピーディスク、CDーROM、Mドディスク、フロッピーディスク、CDーROM、Mに記録され、コンピュータで読み取り可能な記録集件に記録され、コンピュータによって記録媒件から読み出されることによって実行される。またこのプログラムは、上記記録媒件を分して、または伝送媒件として、イは、上記記録媒件を分して、または伝送媒件として、インターネット等のネットワークを介して配布することができる。

【0191】つぎに、実施の形態2~6に係る情報分類接置について説明する。なお、以下説明する実施の形態2~6においては、上記のように多くのノイズを含んだものであるとの解釈に基づいて、一回の文 集合からの話題(内容)抽出と位置づけ、文 分類のためのパラメータ(対象文書集合やクラスタ数、類収度測度、ストップワード等)を変化させながら複数化の分類を実行させ、その結果を保持・統合する手段を設けることで、任意の文書集合にどのような内容が含まれるかを漸次的に収集するものである。

(20)

に係る文書分類装置を構成する情報処理システムは、図 01およびクライアント102のハードウエア構成につ 1に示したように実施の形態1の情報処理システムと同 であるので、その説明は省略する。 いても、図2・図3に示したように実施の形態1と同様 様であるので、その説明は省略する。また、サーバー1 【0192】〔実施の形態2〕この発明の実施の形態2

2による文 分類装置の構成を機能的に示すブロック図 の機能的構成について説明する。図30は、実施の形態 【0193】つぎに、実施の形態2による文書分類装置

08と、分類体系記憶部3009と、クラスタ選択指示 ラスタ特徴表示部3007と、クラスタ特徴算出部30 置は、入力部3001と、言語解析部3002と、ベク 部3010と、分類体系閲覧操作部3011と、を含む トル生成部3003と、分類部3004と、分類パラメ 一夕指示部3005と、分類結果記憶部3006と、ク 【0194】図30のブロック図において、文書分類等

02、RAM203または303、あるいはディスク装 特徴表示部3007、クラスタ特徴算出部3008、分 夕指示部3005、分類結果記憶部3006、クラスタ クトル生成部3003、分類部3004、分類パラメー り、各部の機能を実現する。 U 2 0 1 または301等が命令処理を実行することによ 録されたプログラムに記載された命令にしたがってCP 置306またはハードディスク316等の記録媒体に記 分類体系閲覧操作部3011は、ROM202または3 類体系記憶部3009、クラスタ選択指示部3010、 【0195】入力部3001、書語解析部3002、ペ

文 群を得ることができる1/F204または309年 ナ313、またはネットワーク103を経由して文書や は311、スキャナ313、OCR機能を備えたスキャ 入力するものであり、たとえば、キーボード209また 【0196】ここで、入力部3001は、文書データを

化されている場合に、そのデータベースが記録された媒 体を本実施の形態の文 分類装置に組み入れた場合も文 のすべてを含む。たとえば、文書データがデータベース 【0197】また、入力部3001は、上記以外に、文 データを取得することができるものであれば、それら データの入力とする。

0 1により入力された文書データを解析して言語解析情 前記文書データに対する文書特徴ベクトルを生成するも 解析部3002により得られた言語解析情報に基づいて 報を得るものであり、ベクトル生成部3003は、喜語 【0198】また、書語解析部3002は、入力部30

3003により生成された文 特徴ベクトル間の類似度 【0199】また、分類部3004は、ベクトル生成部

> のであり、分類パラメータ指示部3005は、分類パラ ネットワーク103を経由して指示情報を得ることがで 09または311、マウス210または312、または に基づいて文書を分類し、文書の部分集合を生成するも きる1/F204または309等である。 メータを指示するものであり、たとえば、キーボード2

るものである。また、分類体系記憶部3009は、クラ されたクラスタ特徴を分類体系の構成要素として記憶す 部3009は、クラスタ特徴算出部3008により算出 文書の部分集合に関する情報を記憶するものである。ま 3004により分類された結果、すなわち、分類された 選択するものである。また、分類体系閲覧操作部301 り生成された文書の部分集合の中から所望の部分集合を 表示部3007により表示された複数のクラスタ特徴の クラスタに所属する全ての文書もしくは所属する文書の なわち、クラスタ選択指示部3010により選択された 合を分類体系の構成要素として記憶するものである。す スタ選択指示部3010により選択された文書の部分集 ラスタ特徴を算出するものである。また、分類体系記憶 004により生成された文書の部分集合の特徴であるク 【0201】クラスタ特徴算出部3008は、分類部3 **部3008により算出されたクラスタ特徴を表示する。** た、クラスタ特徴表示部3007は、クラスタ特徴算出 ラスタ選択指示部3010は、前記分類部3004によ 中から所望のクラスタを選択するものである。また、 【0200】また、分類結果記憶部3006は、分類的 覧したい場合に、その閲覧の操作をおこなうものであ 一部を分類体系の構成要素として記憶するものである。 1は、分類体系記憶部3009に記憶されたデータを関 【0202】クラスタ選択指示部3010は、クラスタ

査等により得られた自由記述回答の分析場面を想定し、 を抽出することが重要となる好適な例を、アンケート調 その具体例を用いて説明する。 【0203】つぎに、文書集合に含まれる話題(内容)

が可能であり、このような機能を用いて大量のテキスト て短期間に数千~数万件の自由記述回答を回収すること 【0204】近年、たとえば、インターネット等を介し

スト情報の収集の例として、「オフィスのネットワーク 書で答えた一つの回答記述を文書とすると、文書集合 化による無駄を挙げてください」という質問に対して文 【0205】アンケート調査により得られた大量のテキ (クラスタ) は1件ごとの回答の集合ということにな

は、そのニーズの一つとして、意見集合(文書集合)に すべく、話題の抽出を類似する意見のまとまり(分類) 概略を把握したい場合がある。このようなニーズを満た どのような種類の意見 (話題) が含まれており、意見の 【0206】ここで、操作者(アンケートの分析者)

情報の収集をおこなうことができる。

が含まれているかを抽出する。 により実現し、アンケート結果にどのような種類の意見 【0207】文書分類は、典型的には大きく分けてつき

は、入力部3001により入力された各文書(意見)に の3段階のステップから構成される。第1ステップで 際、たとえば、形態素形跡等の言語解析アルゴリズムが ついて、書語解析部3002が、各文書に含まれる単語 し、各文書を行とし、要素を単語の出現頻度とした「単 【0208】第2ステップでは、抽出された単語を列と (あるいは、特定の連続する文字列)を抽出する。この

用いると単語抽出のほかに、単語の品詞情報、複合語 き、こうした情報を上記単語×文書の行列を生成する 態素解析機能と構文解析機能を有する書語解析ツールを 語」×「文書」の行列が生成される。なお、一般的な形 (フレーズ)、構文情報等の同時に取得することがで

間内に各文書をベクトル表現する。これには、以下の方 ×「文書」の行列に基づいて単語で構成される多次元空 装している。 法があり、本実施の形態においては、すべての方法を実 【0209】ベクトル生成部3003は、この「単語」

際、考慮することができる。

て値の重み付けをする方法、(3)上記行列から文書間 類対象全体の文書集合内での各単語の出現頻度を考慮し こなわれる)を適用して潜在的意味空間を構成する方 子分析や主成分分析、数量化理論第3類等を利用してお の内積行列を算出し、これに特異値分解(たとえば、因 法、(2)各文書の長さ(文字の数やページ数等)や分 【0210】(1)行列の列成分をそのまま利用する方

記潜在的意味空間への変換手法を一般化し、文書間の内 Science, ページ: 254-271, Vol. 4 名:Journal of the American Documents Usingan Explici の喪現空間変換関数を導出しているものもあり、この方 らの類似性を反映する空間へ文書や単語を射影するため 積行列に、文書が有するほかの文書への参照情報から生 6 No. 4. 発行年:1995)」においては、上 Science, 学会名: the American 成される共参照情報などを付加した行列を用いて、これ and Richard K. Beiew. 路文 ties (著者名:Brian T. Bartel I. Garrison W. t Model of Their Similar 【0211】また、「Representating Society for Information Society for Information Cottrell.

【0212】第3ステップでは、分類部3004が、文 特徴ペクトルの類似度を用いて女 を分類する。具体

法も利用することができる。

分析の方法、クラスタリングの方法等を適用することに 的には分類対象データに対してカイ自乗法の手法、判別 より分類が実行される。

施の形態においては、いずれの方法を用いてもよい。 クリッド距離、マハラノピスの距離等が考えられ、本実 いずれの方法を用いてもよい。 スタリングが考えられるが、本実施の形態においては、 してもさまざまなものが公知になっている。クラスタリ ングは、大別して階層型クラスタリングと非階層型クラ 【0214】また、クラスタリングのアルゴリズムに関 【0213】また、類似度としては、内積や余弦、ユー

類パラメータを指示する。分類部3004は、分類パラ にしたがって内部に保持される文 特徴ベクトルを分類 分類部3004が文書特徴ペクトルを分類するための分 メータ指示部3005により指示された分類パラメータ 【0215】また、分類パラメータ指示部3005は、

が終了すると、分類結果は分類結果記憶部3006によ ップの各処理を実行することにより第1回目の文 分類 り保持される。 【0216】このようにして、第1ステップ~第3ステ

のかを示す特徴、すなわちクラス特徴を算出する。典型 に基づいて文書をソーティングして出力する。 一部を算出するが、その際、クラスタの重心との類似度 的には各クラスタに所属する文 、あるいはその文 の が、分類結果がどのようなクラスタを得ることができた 【0217】引き続き、クラスタ特徴算出部3008

の程度を表すクラスタ内の標準偏差のような数値をクラ スタの特徴を表現するものとして算出する。 スタに所属する文書数、クラスタ内での文 のばらつき 【0218】そのほか、クラスタ内で最頻の単語、クラ

を示すものであれば、上配の内容(特徴)以外のもので 算出されるものであり、操作者に対してクラスタの特徴 対して出力(表示)されたクラスタがどのようなもの 【0219】これらのクラスタの特徴情報は、操作者に (どのような特徴を有するもの) かを把握させるために

型クラスタリングの場合は、クラスタ重心間の類似度に の場合は、その上位あるいは下位のクラスタを、非階層 間の関係を示す情報も算出する。階層型クラスタリング 配のようにクラスタの特徴を示すもの以外に、クラスタ **基厶へ近接のクラスタを算出する。** 【0220】また、クラスタ特徴算出部3008は、上

るクラスタ特徴の表示およびクラスタ選択の内容につい 置のクラスタ特徴表示部3007の表示の一例を示す説 て説明する。図31は、実施の形態2による文 分類装 【0221】つぎに、クラスタ特徴表示部3007によ

できるようになっており、各クラスタは「クラスタ1 【0222】図31において、クラスタ単位で操作者が

D」欄3101、「メンバー数」欄3102、「頻度の 心との類似度」欄3105等の項目から構成される。 高い単語」禰3103、「文書内容」禰3104、「重 夕のIDを示す番号が通し番号で付与され、表示され 【0223】「クラスタID」概3101には、クラス 「メンバー数」欄3102はクラスタに所属する文

の中で頻度の高い単語が抽出され「頻度の高い単語」標 3 1 0 3 に表示される。「文書内容」欄 3 1 0 4 には文 は、数値化された重心との類似度が表示される。これに より、操作者の理解容易性が向上する。 の内容が表示され、「重心との類似度」欄3105に あるいは文 の一部の数が算出され、表示される。そ

る。ここで、内容(特徴)が理解可能なクラスタが一つ りクラスタを選択することができる。 でもあれば、操作者はクラスタ選択指示部3010によ **少いてクラスタについてその特徴を把握することができ** 【0224】操作者は、表示された情報(特徴量)に基

れるようにしてもよい。 てが選択されるわけではなく、その一部の文書が選択さ る。なお、選択したクラスタに所属する文書は必ずすべ 当該クラスタIDのクラスタ全体を選択することができ 10を移動させ、その位置でクリックすることにより、 たとえば、「クラスタID」 概3101ヘカーソル31 2 等によって、表示されているクラスタの所定の位置、 【0225】より具体的には、マウス210または31

転表示しており、当該クラスタ(クラスタ1D「1」) 101がクリックされ、これにより、クラスタ全体が反 が選択されたことを示している。 **【0226】図31においては、「クラスタID」 欄3**

05により分類パラメータの再設定をおこない、再度分 類実行をおこなうことができる。 ラスタが存在しない場合は、分類パラメータ指示部30 【0227】また、操作者は、内容が理解可能であるク

部3008からクラスタに関する上記特徴量を検索し記 09へ送信される。分類体系記憶部3009は、このク れたクラスタIDに関するデータは分類体系記憶部30 ラスタIDに関するデータに基づいてクラスタ特徴算出 【0228】クラスタ選択指示部3010により選択さ

り入力されたクラスタに関するコメント(たとえば、 する。さらに、分類体系記憶部3009は、操作者によ に、分類結果記憶部3006から分類結果を検索し記憶 【0229】また、分類体系記憶部3009は、同様

類体系の利用価値がより向上する。 報を分類体系の構成要素として記憶することにより、分 憶することもできる。このように、操作者が作成した情 「ネットワークの維持費が高い」等)の情報を併せて記

設けることにより、選択・保持するクラスタの内容の関 されたデータは、別途閲覧操作用のインターフェイスを 【0230】なお、分類体系記憶部3009により記憶

> 持されているクラスタ重心間の類似度等を用いて自動 覧や、クラスタ間の意味的な関連を手動であるいは、保 で、構造化・体系化することができる。

まず、分類の対象となる文書が入力される(ステップS ーチャートである。図32のフローチャートにおいて、 態2による文書分類装置の一連の処理の手順を示すプロ 連の処理の手順について説明する。図32は、実施の形 【0231】つぎに、実施の形態2の文書分類装置の一

れる (ステップS3203)。 抽出された単語に基づいて、文書特徴ベクトルが生成さ 【0232】つぎに、入力された文書の言語が解析され (ステップS3202)、解析された結果、すなわち、

果、すなわち、クラスタに関する情報を記憶する(ステ たがって文書を分類し(ステップS3205)、その結 S3204肯定)は、指示があった分類パラメータにし 待って、分類パラメータの指示があった場合(ステップ 【0233】その後、分類パラメータの指示があるのを ップS3206)。

ラメータの指示があるのを待つ(ステップS320 ら、クラスタが選択されたか否かを判断し(ステップS 9否定) は、ステップS3204へ移行し、再度分類パ 3209)、選択されなかった場合(ステップS320 し(ステップS3207)、算出された結果を表示する 【0234】つぎに、分類されたクラスタの特徴を算出 (ステップS3208)。 表示されたクラスタの中か

る(ステップS3210)。この際、操作者により入力 選択されたクラスタに関して分類体系を生成し、記憶す スタが選択された場合(ステップS3209肯定)は、 されたクラスタに関する情報を併せて記憶することもで きる。これにより、一連の処理を終了する。 [0235] 一方、ステップS3209において、クラ

間変換関数を算出し、その表現空間で文書分類をおこな **な関連性を反映しうる表現空間へ変換するための表現空** 文書分類装置によれば、分類対象である文書群での文書 間の類似性に基づいて、各文書をそれら文書間の意味的 現することができる。 うことにより、操作者の意図を反映しうる文書分類を実 【0236】以上説明したように、実施の形態2による

間の類似度等を用いて、クラスタの内容に基づくクラス タを得ることができるとともに、クラスタ特徴算出部3 タの構造化・体系化をおこなうことができる。 008・分類体系記憶部3009により、クラスタ重心 【0237】したがって、分類部3004によりクラス

選択されたクラスタのみを用いて、より操作者の意図し たものに近いクラスタの構造化・体系化をおこなうこと 【0238】また、クラスタ選択指示部3010により

【0239】〔実施の形態3〕さて、上述した実施の形

憩2に加えて、以下に説明する実施の形態3のように、 さらにベクトル記憶部と、ベクトル修正部とを含む構成 とするようにしてもよい。

バー101およびクライアント102のハードウエア袋 成についても、図2・図3に示したように実施の形態 1と同様であるので、その説明は雀略する。また、サー る情報処理システムは、図1に示したように実施の形態 と同様であるので、その説明は省略する。

実施の形態3による文書分類装置の構成を機能的に示す の機能的構成について説明する。図33は、この発明の 30と同一のものに関しては同じ符号を付して、その説 ブロック図である。図33において、実施の形態2の図 【0241】つぎに、実施の形態3による文書分類装置

部3005、分類結果記憶部3006、クラスタ特徴表 置は、入力部3001、言語解析部3002、ベクトル と、ベクトル修正部3302とを含む構成である。 系閲覧操作部3011のほか、ベクトル記憶部3301 記憶部3009、クラスタ選択指示部3010、分類体 示部3007、クラスタ特徴算出部3008、分類体系 生成部3003、分類部3004、分類パラメータ指示

たのこりとなるように修正するものである。 れた部分集合に属する文書の文書特徴ベクトルを除去し ベクトルを、クラスタ選択指示部3010により選択さ 特徴ペクトル記憶部3301により記憶された文書特徴 るものである。また、ベクトル修正部3302は、文書 部3003により生成された文書特徴ペクトルを記憶す 【0243】ベクトル記憶部3301は、ベクトル生成

3302により修正された文書特徴ベクトルに基づいて [0244] また、分類部3004は、ベクトル修正部

はハードディスク316等の記録媒体に記録されたプロ **修正部3302は、ROM202または302、RAM** は301等が命令処理を実行することにより、各部の機 203または303、あるいはディスク装置306また グラムに記載された命令にしたがってCPU201また

ベクトル(行ベクトル)はベクトル記憶部3301によ た文書特徴ベクトル(列ベクトル)・単語(単語特徴) って記憶される。これは、次回以降の分類実行の際に利 【0246】ベクトル生成部3003において生成され

特徴ベクトルはベクトル記憶部3301により記憶され これらの文書が除かれるよう削除する。削除された文書 書のすべてあるいはその一部の文書を除き、次回以降も 指示部3010により選択されたクラスタに所属する文 【0247】ベクトル修正部3302は、クラスタ選択

【0240】実施の形態3による文書分類装置を構成す

【0242】図33のブロック図において、文書分類装

文書を分類する。

【0245】なお、ベクトル記憶部3301、ベクトル

用する文書特徴ベクトルを確保するためである。

に所属する文書(もしくは操作者に指定されたその一 されているベクトルデータのうち、選択されたクラスタ 【0248】この結果、ベクトル記憶部3301に記憶

部)列ベクトルを除いたものが、次回以降の分類が実行 態3よる文書分類装置の一連の処理の手順を示すフロー 連の処理の手順について説明する。図34は、実施の形 される際に利用されるデータとなる。 【0249】つぎに、実施の形態3の文 分類装置の一 チャートである。図2のフローチャートにおいて、ま

ず、分類の対象となる文 が入力される (ステップS3

ルが記憶される(ステップS3404)。 抽出された単語に基づいて、文 特徴ベクトルが生成さ れ(ステップS3403)、生成された文 特徴ベクト (ステップS3402)、解析された結果、すなわち、 【0250】つぎに、入力された文 の言語が解析され

果、すなわち、クラスタに関する情報を記憶する(ステ S3405肯定)は、指示があった分類パラメータにしたがって文書を分類し(ステップS3406)、その結 待って、分類パラメータの指示があった場合(ステップ 【0251】その後、分類パラメータの指示があるのを ップS3407)。

ラメータの指示があるのを待つ(ステップS340 O否定)は、ステップS3405へ移行し、再度分類パ 3410)、選択されなかった場合(ステップS341 ら、クラスタが選択されたか否かを判断し(ステップS し(ステップS3408)、算出された結果を表示する 【0252】つぎに、分類されたクラスタの特徴を算出 (ステップS3409)。 表示されたクラスタの中か

る(ステップ3411)。この際、操作者により入力さ 選択されたクラスタに関して分類体系を生成し、記憶す スタが選択された場合(ステップS3410肯定)は、 か否かを判断する(ステップS3412)。 る。その後、繰り返し処理をおこなう旨の指示があった れたクラスタに関する情報を併せて記憶することもでき 【0253】一方、ステップS3410において、クラ

S3413の各処理を繰り返しおこなう。 べてあるいはその一部の文 を除くように文 特徴ベク ップS3405へ移行し、以後、ステップS3405~ トルを修正する(ステップS3413)。その後、ステ 12肯定)は、選択されたクラスタに所属する文 のす 処理をおこなう旨の指示があった場合(ステップS34 【0254】ステップS3412において、繰り返して

3 4 1 2 否定)は、これにより、一連の処理をすべて終 返して処理をおこなう旨の指示がない場合(ステップS 【0255】一方、ステップS3412において、繰り

【0256】以上説明したように、実施の形態3による

特開2000-285140

24)

文 分類装置によれば、ベクトル修正部3301により、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0257】(実施の形態4】さて、上述した実施の形態3においては、ベクトル配憶部およびベクトル棒圧部とを含む構成であったが、以下に説明する実施の形態4のように、ベクトル棒圧部に代わりに、文書表現空間棒正部を含む構成とするようにしてもよい。

[0258] 実施の形態4による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウエア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0259】つぎに、実施の形態4による文書分類装置の機能的構成について説明する。図35は、この発明の実施の形態4による文 分類装置の構成を機能的に示すプロック図である。図35において、実施の形態2の図30と同一のものに関しては同じ符号を付して、その説明を省略する。

【0262】また、分類部3004は、文書表現空間棒正部3502により棒正された文書表現空間を用いて、ベクトル生成部3003により生成された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【0263】なお、ベクトル記憶部3501、文書表現空間修正部3502は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがつてCPU201または301等が命令処理を実行することにより、各部の地能を取刊する。

【0264】つぎに、文 表現空間棒正部3502の内 容について説明する。実施の形態3におけるベクトル棒 正部3302にあっては、既知になったクラスタの影響

を排除するために文書特徴ペクトルを除去するが、文書特徴ペクトルを表現する多次元空間自体の変更はおこなわれない。

【0265】したがって、前回の分類実行の結果、操作者により選択されたクラスタの形成特徴を次回の分類実行の際に排除したい場合は、文書ベクトルを表現する空間自体の変更が必要となる。

【0266】そこで、文書表現空間修正約3502を備え、文書表現空間の修正をおこなうものである。ここで、文書表現空間の特徴次元を変更する例として、操作で、文書表現空間の特徴次元を変更する例として、操作者により選択されたクラスタの重心と類似度の高い特徴次元の削除をおこなうことについて説明する。

【0267】操作者により選択されたクラスタの重心は ベクトルとして表現することができるので、このクラス タ重心ベクトルとベクトル記憶部3501に記憶されて いる文書表現空間の各特徴次元との類似度を算出することにより、類似度の高い特徴次元を判別する。

【0268】なお、類似の測度としては、余弦、内積、ユーケリッド距離、マハラノビス距離等を用いる。また、判別に関してはある類似度以上を削除対象として採用するようなしきい値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理による判別を用いる。また、判別分析等も用いることができる。

【0269】文書表現空間修正的3502は、上述のような削除対象の特徴次元を算出して、特徴次元の削除をおこなう。特徴次元の削除は、ベクトル記憶的3501に記憶されている「特徴次元(単語)」×「文書」の行列から判別された特徴次元について行ベクトルを削除することによりおこなう。文書表現空間修正部3502により毎正された文書ベクトルは、次回以降の分類のために、ベクトル記憶的3501に記憶される。

【0270】つぎに、実施の形態4の文書分類装置の一連の処理の手順について説明する。図36は、実施の形態4よる文書分類装置の一連の処理の手順を示すフローチャートである。図36のフローチャートにおいて、まず、分類の対象となる文書が入力される(ステップS3601)。

【0271】つぎに、入力された文書の言語が解析され(ステップ53602)、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され(ステップ53603)、生成された文書特徴ベクトルが記憶される(ステップ53604)。

【0272】その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合(ステップS3605肯定)は、指示があった分類パラメータにしたがって文書を分類し(ステップS3606)、その結果、すなわち、クラスタに関する情報を記憶する(ステップS3607)。

【0273】つぎに、分類されたクラスタの特徴を算出

し(ステップS3608)、算出された結果を表示する(ステップS3609)。表示されたクラスタの中から、クラスタが選択されたか否かを判断し(ステップS3610)、選択されなかった場合(ステップS3610 で否定)は、ステップS3605へ移行し、再度分類パラメータの指示があるのを待つ(ステップS360

【0274】一方、ステップS3610において、クラスタが選択された場合(ステップS3610肯定)は、選択されたクラスタに関して分類体系を生成し、記憶する(ステップ3611)。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する(ステップS3612)。

[0275] ステップS3612において、繰り返して 処理をおこなう旨の指示があった場合(ステップS36 12情報)」×「文章」の行列 から判別された特徴次元について行ベクトルを削除する ことにより文書表現空間を修正する(ステップS3613の各処理を繰り返しおこ テップS3605~S3613の各処理を繰り返しおこか 3)。その後、ステップS3605~S3613の各処理を繰り返しおこか 3

【0276】一方、ステップS3612において、繰り返して処理をおこなう旨の指示がなかった場合(ステップS3612否定)は、これにより、一連の処理を終了する。

【0277】以上説明したように、実施の形態4による文書分類装置によれば、前回の分類実行の結果、文書表現空間修正部3502により操作者に選択されたクラス現空間修正部3502により操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0278】 (実施の形態5】さて、上述した実施の形態3または実施の形態4においては、ベクトル棒正虧または文書表現空間棒正部のいずれか一方のみを含む構成であったが、以下に説明する実施の形態5のように、ベクトル棒正部および文書表現空間棒正部の両方を含む構成とするようにしてもよい。

【0279】実施の形態5による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウエア様成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0280】つぎに、実施の形態5による文書分類装置の機能的構成について説明する。図37は、この発明の実施の形態5による文書分類装置の構成を機能的に示すブロック図である。図37において、実施の形態2の図30と同一のものに関しては同じ符号を付して、その訳明を失いする。

【0281】図37のプロック図において、文_分類装置は、入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴其出部3008、分類体系記憶部3009、クラスタ透校指示部3010、分類体系記憶部3009、クラスタ透校指示部3010、分類体系別質操作部3011のほか、ベクトル記憶部3701と、ベクトル接正部3702と、文 表現空間棒正部3703とを含む構成である。

【0282】ベクトル記憶部3701は、ベクトル生成部3003により生成された文 特徴ベクトルを記憶するものである。また、ベクトル棒正部3702は、文特徴ベクトル記憶部3701により記憶された文 特徴ベクトルを分類部3004により生成された文 の部分集合の文書特徴ベクトルを除去したのごりの文 特徴ベクトルをなるように修正するものである。

【0283】また、文 表現空間停圧部3703は、ベクトル配億部3701により配億された文 特徴ベクトル間の類似度を判断する際の文 表現空間を前記クラスタ選択指示部3010により選択されたクラスタ特徴に 基づいて停正するものである。

【0284】また、分類約3004は、文 表現空間棒正約3703により棒正された文書表現空間を用いて、ベクトル棒正約3702により棒正された文書特徴ベクトル間の類似度に基づいて文 を分類する。

【0285】なお、ベクトル記憶部3701、ベクトル棒正部3702、文書表現空間棒正部3703は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0286】つぎに、ベクトル修正約3702および文書表現空間修正約3703の内容について説明する。実書表現空間修正約3703の内容について説明する。実施の形態4においては、選択されたクラスタに所属する文書は次回以降の分類実行の際にも使用される。

【0287】実施の形態5では、ベクトル棒正約3702および文書表現空間棒正約3703の両方を具備することにより、選択されたクラスタに所属する文を次回の分類実行の際に除去し、次回の分類実行の際には分類対象文書としないようにする。

【0288】実施の形態4においては、話題抽出の側面を強調し、ある文書が複数の話題として分類される可能性を前提としており、たとえば、ネットワーク化に関する調査における「エンドユーザーがソフトウエアのインストール方法について聞いてくるのでシステム管理者としての仕事ができない」という回答について言えば、この意見は「ソフトウエアの操作方法理解に関する困難性」という話題として分類される可能性もあきの仕事の多忙さ」という話題で分類される可能性もあ

(26)

うクラスタの両方とも抽出したいというニーズに応えて いうクラスタと「システム管理者の仕事の多忙さ」とい も、「ソフトウエアの操作方法理解に関する困難性」と 【0289】実施の形態4においては、いずれにして

分類を実行する際、分類対象から除去するものである。 正部3702により、n回目の分類で選択されたクラス 形態5では、このような要求に応えるため、ベクトル修 話題は既知であるので、次回の分類の際はなるべく異な クトル修正部3702では劣ベクトルを除去すること 1において列ベクトルの形式で記憶されているため、ベ **示を受けたクラスタの所属文書はベクトル記憶部370** タに所属する文 のすべてまたはその一部を次回以降の る分類結果が欲しいとするケースも考えられる。実施の で、次回以降の分類実行用の分類対象文書集合を生成す 【0291】クラスタ選択指示部3010により選択指 【0290】これとは反対に、操作者は、一度抽出した

トル記憶部3701に記憶されている行列から特徴次元 たクラスタにより文 表現空間修正部3703は、ベク 【0292】さらに、実施の形態4と同様に、選択され

ず、分類の対象となる文 が入力される (ステップS3 チャートである。図38のフローチャートにおいて、ま 態5よる文 分類装置の一連の処理の手順を示すフロー 連の処理の手順について説明する。図38は、実施の形 【0293】つぎに、実施の形態5の文書分類装置の一

れ(ステップS3803)、生成された文書特徴ベクト 抽出された単語に基づいて、文書特徴ベクトルが生成さ ルが記憶される(ステップS3804)。 (ステップS3802)、解析された結果、すなわち、 【0294】つぎに、入力された文書の言語が解析され

待って、分類パラメータの指示があった場合(ステップ 果、すなわち、クラスタに関する情報を記憶する(ステ S3805肯定)は、指示があった分類パラメータにし ップS3807)。 たがって文 を分類し(ステップS3806)、その結 【0295】その後、分類パラメータの指示があるのを

し(ステップS3808)、算出された結果を表示する ラメータの指示があるのを待つ(ステップS380 0否定)は、ステップS3805へ移行し、再度分類パ 3810)、選択されなかった場合(ステップS381 ら、クラスタが選択されたか否かを判断し(ステップS (ステップS3809)。 表示されたクラスタの中か 【0296】つぎに、分類されたクラスタの特徴を算出

スタが選択された場合(ステップS3810肯定)は. [0297] 一方、ステップS3810において、クラ

> る。その後、繰り返し処理をおこなう旨の指示があった れたクラスタに関する情報を併せて記憶することもでき 選択されたクラスタに関して分類体系を生成し、記憶す る (ステップ3811)。この際、操作者により入力さ か否かを判断する(ステップS3812)。

べてあるいはその一部の文書を除くように文書特徴べク 処理をおこなう旨の指示があった場合(ステップS38 トルを修正する(ステップS3813)。 【0298】ステップS3812において、繰り返して 2 肯定)は、選択されたクラスタに所属する文書のす

元(単語)」×「文書」の行列から判別された特徴次元 S3805へ移行し、以後、ステップS3805∼S3 814を繰り返しおこなう。 を修正する(ステップS3814)。その後、ステップ について行べクトルを削除することにより文書表現空間 【0299】ステップS3813に引き続き、「特徴次

返して処理をおこなう旨に指示がない場合(ステップS 3812否定)は、これにより、一連の処理をすべて終 【0300】一方、ステップS3812において、繰り

修正部3703が、前回の分類実行の結果、操作者に選 することができ、排除した状態で新たなクラスタを生成 になったクラスタの影響を排除し、かつ、文書表現空間 することができる。 択されたクラスタの形成特徴を次回の分類実行時に排除 書分類装置によれば、ベクトル修正部3702が、既知 【0301】以上説明したように、実施の形態5よる文

選択情報をクラスタ特徴とともに表示するようにしても 憩2または実施の形態4においては、繰り返し分類処理 施の形態6のように、選択情報付与部を含む構成とし、 情報については考慮していなかったが以下に説明する実 をおこなった場合に、ある文書が何度選択されたかその 【0302】〔実施の形態6〕さて、上述した実施の形

と同様であるので、その説明は省略する。 成についても、図2・図3に示したように実施の形態1 バー101およびクライアント102のハードウエア賞 る情報処理システムは、図1に示したように実施の形態 【0303】実施の形態6による文書分類装置を構成す 1と同様であるので、その説明は省略する。また、サー

実施の形態6による文書分類装置の構成を機能的に示す 35と同一のものに関しては同じ符号を付して、その説 ブロック図である。図39において、実施の形態4の図 の機能的構成について説明する。図39は、この発明の 【0304】つぎに、実施の形態6による文書分類装置

生成部3003、分類部3004、分類パラメータ指示 部3005、分類結果記憶部3006、クラスタ特徴表 置は、入力部3001、言語解析部3002、ベクトル **【0305】図39のブロック図において、文書分類装**

> 表現空間修正部3502のほか、選択情報付与部390 系閲覧操作部3011、ベクトル記憶部3501、文書 記憶部3009、クラスタ選択指示部3010、分類体 示部3007、クラスタ特徴算出部3008、分類体系 1を含む構成である。

の記録媒体に記録されたプログラムに記載された命令に 202または302、RAM203または303、ある 付与部3901により付与された選択情報を表示する。 007は、クラスタ特徴を表示するとともに、選択情報 示す選択情報を付与する。また、クラスタ特徴表示部3 べてあるいは一部が選択された場合に選択されたことを 4により生成された文書の部分集合に所属する文書のす 行することにより、機能を実現する。 したがってCPU201または301等が命令処理を実 いはディスク装置306またはハードディスク316等 【0306】選択情報付与部3901は、分類部300 【0307】なお、選択情報付与部3901は、ROM

内容について説明する。アンケートの調査の例におい 予想できなかった意見である場合が多いからである。 験的に知られている。これは、調査を企画する担当者が て、独自性の高いユニークな意見は貴重であることが経 【0308】つぎに、選択情報付与部3901の詳細な

度も選択されない文書の識別性も向上させることができ において、クラスタ特徴表示部3007で個々の文書を で、多重に利用される文書の識別性を向上させ、かつ-表示する際に、各文書が何回選択されたかを示すこと 属する文書を、次回以降の分類実行の際に使用する場合 【0309】そこで、操作者に選択されたクラスタに所

は、各文書がどのサイクルに分類実行の際に操作者に選 選択情報として「0」を記録する。 択情報として「1」を記録し、選択されなかった場合に 俄の分類結果記憶部3006において設けられたテープ 択されたかを記録する。すなわち、選択された場合は選 1 ロごとにテーブル化されており、テーブル4000 ル4000を示す説明図である。図40において、文書 【0310】図40は、実施の形態6による文書分類装

方、文書IDの「2」は、未だ一度も選択されておら ず、操作者にとって未知の意見という可能性を示唆して 分類実行時には選択されなかったことを示している。— 操作者に選択されたことを示し、第3回目、第4回目の I Dの「1」、第1回目および第2回目の分類実行時に 【0311】たとえば、4回分類が実行された際、文書

背景の色の濃度や彩度等が考えられる。 よい。変化させる視覚的特性としては、たとえば文字や 選択された回数に応じて表示を変化させるようにすると 示部3007が文書を操作者に表示する際、たとえば、 【0312】こうした情報に基づいて、クラスタ特徴表

【0313】また、直接的に数字やグラフ等で選択され

嫌煩できる表示形式であれば、上記のものに限らない。 3011の閲覧操作により閲覧できるようにしてもよ される文書と一度も選択されていない文 とを視覚的に た回数を表現することもできる。いずれにしてもよ選択 【0314】また、上記選択情報を分類体系閲覧操作部

テップS 4 1 0 2)。 01)、それに引き続き、最初の文 が抽出される (ス おいて、まず、分類処理がおこなわれ(ステップS41 示すフローチャートである。 図4 1のフローチャートに 文書分類装置の選択情報付与部3901の処理の手順を 内容について説明する。図41は、実箔の形態6による 【0315】つぎに、選択情報付与部3901の処理の

おける分類処理の際に選択されたか否かを判断する(ス で、すべての文書について処理が終了していない場合 てデータ「0」を記録する(ステップS4105)。 記録する(ステップS4104)。一方、選択されなか テップS4103)。ここで、選択された場合(ステッ し、以後、ステップS4103~S4107を繰り返し したか否かを判断する(ステップS 4 1 0 6)。ここ った場合(ステップS4103否定)は、選択情報とし プS4103肯定)は、選択情報としてデータ「1」を (ステップS4107) 、ステップS4103へ移行 (ステップS4106否定) は、つぎに文書を抽出し 【0317】つぎに、すべての文(について処理が終了 【0316】抽出された文 が、ステップS4101に

処理がおこなわれる(ステップS4101)。このようにして、分類処理がおこなわれる回教だけ、ステップS 06肯定)は、ステップS4101へ移行し、再度分類 ての文書について処理が終了した場合(ステップS41 4 1 0 1~S 4 1 0 7 の各処理が繰り返しおこなわれ 【0318】一方、ステップS4106において、すべ

ば、選択情報付与部3901が選択情報を付与し、その れない文書の識別性を向上させることができる。 で、多重に利用される文 の鎌別性および一度も選択さ 選択情報をクラスタ特徴表示部3007が表示するの 【0319】以上説明したように、実施の形態6によれ **【0320】なお、実施の形態2~5で説明した文 分**

で実行することにより実現される。このプログラムは、 類方法は、あらかじめ用意されたプログラムをパーソナ ンターネット等のネットワークを介して配布することが は、上記記録媒体を介して、または伝送媒体として、イ 出されることによって実行される。またこのプログラム 体に記録され、コンピュータによって記録媒体から読み MO、DVD等のコンピュータで読み取り可能な記録娘 ハードディスク、フロッピーディスク、CD-ROM. **ルコンピュータやワークステーション等のコンピュータ**

(28)

文化・科学技術等の特定分野に分類される新聞記事も文 分類等により分類される公開特許公報や、政治・経済・ 合、それを文 と言う。具体的な例をあげれば、IPC その一つ以上の文の集まりが分類される対象である場 然言語で記述された一つ以上の文の集まりであり、かつ 類装置について説明する。本発明の実施の形態では、自 【0321】つぎに、実施の形態7~16に係る情報分 であるし、それらから請求項や特定の一文を取り出し

図42に示したように、実施の形態7の文書分類装置 の形態7を示す文 分類装置の構成プロック図である。 たものであっても、請求項という分類に含まれる文であ とみなす。以下、図面によりこの発明の実施の形態1~ るか、用途等により分類可能な特定の一文であれば文書 【0322】〔実施の形態7〕図42はこの発明の実施

段)5001、それぞれの文書データを所定の基準に基 は、文 データ群を入力する文書入力部(文書入力手 成手段)5003を備えている。 割文 対応マップ生成部(文書一分割文書対応マップ生 割文 データとを対応付けるマップを生成する文書一分 分割部(文 分割手段)5002、上記文書データと分 **づいて一つまたは複数の分割文書データに分割する文書** 【0323】また、上記文書分類装置は、分割文書デー

手段)5005、上記文 一分割文書対応マップと上記 成する分割文 分類結果生成部 (分割文書分類結果生成 タつまり分割された文 を分類する分割文書分類部(分 生成手段)5006などを備えている。 結果情報を生成する文 分類結果生成部 (文書分類結果 分割文 分類結果情報とを用いて上記文書データの分祭 割文 分類手段)5004、分割文書分類結果情報を生

生成部5006は共有または独自のプログラム記憶用メ モリおよびプログラムにしたがって動作するCPUを有 04、分割文 分類結果生成部5005、文書分類結果 割文 対応マップ生成部5003、分割文書分類部50 [0324] なお、上記文書分割部5002、文書一分

態7の文 分類装置、文 分類方法を詳細に説明する。 まず、文 入力部5001により、文書群が入力され え、それらのいずれか一つを介して文書データ群を入力 置、着脱型記録媒体、またはネットワーク通信手段を係 る。上記文 入力部5001はキーボード、OCR装 【0325】以下、図42などにしたがって、実施の形

する方法としては、文 の構造情報や文書を構成する要 の分割文 データを生成する。なお、文書データを分割 基づいて分割し、一つの文書データから一つまたは複数 ータ群を取得し、それぞれの文書データを所定の基準に が、ここでは、その方法は問わないこととする。 素情報を用いたり、利用者が指定する方法などを用いる 【0326】そして、文 分割部5002が上記文書デ

> により分離されているので、この規則を用いて一つの文 タとして含めることもできるが、ここでは含めないこと 一夕を生成する。なお、分割前の文書1も分割文書デー 形成される分割文書1-1~1-7の7つの分割文書デ ックが文書単位となっている。図示したように、この文 数のニューストピックが記述されており、1日分のトピ 夕を生成する一例を示す。この例に示した文書1には複 法でおこなわれる、文書データから複数の分割文書デー 書である文書1を分割し、一つが一つのトピックにより 書ではそれぞれのニューストピックが二つの改行コード 【0327】図43に、この文書分類装置/文書分類方

成する一例を示す。図44において、文書1~文書3は 応付ける方法についてはここでは問わないこととする。 するのである。なお、文書データと分割文書データを対 割文書データを一意に示す識別子からなるマップを生成 とから構成されるマップ、あるいは文書データごとに分 示す識別子と個々の分割文書データを一意に示す識別子 ップを生成する。たとえば、個々の文書データを一意に マップ生成部5003が分割前の文書データとその文書 れらの識別番号を同一にしてもよい。 れる基準において同一とみなすことができる場合は、そ お、任意の複数の分割文書データが文書分類にて用いら データの識別番号と分割文書データの識別番号とを図4 文書データを示し、分割文書1~分割文書12は分割文 データから生成された分割文書データとを対応付けるマ 4の左下に示したテーブル形式で対応づけている。な ることかできる識別番号(雛別子)を付与し、上記文書 データおよび分割文書データにそれぞれを一意に識別す 書データを示している。図示のように、それぞれの文書 【0329】図44に、文書一分割文書対応マップを生 【0328】文書が分割されると、文書一分割文書対応

てそれぞれの文書の特徴を計量的に表す特徴ベクトルを いるそれぞれの単語の出現頻度を計数し、それに基づい 対して、たとえば、言語処理を施し、文書中に含まれて 割文書を対象に文書分類をおこなう。個々の分割文書に により文書分類をおこなう。 分析手法、またはクラスタ分析手法などを適用すること 求め、それらの特徴ベクトルに対してカイ自乗法、判別 【0330】続いて、分割文書分類部5004が上記分

づいた分割文書分類結果情報を生成する。 類結果生成部5005が上記の分割文書分類の結果に基 【0331】つぎに、図45に示すように、分割文書分

えば、図45に示した「分類カテゴリに関する情報」と 報)、生成された所属カテゴリ個々に関する情報(たと り」および「所属カテゴリの代表値との距離」の項の情 カテゴリに分類した結果」という表中の「分類カテゴ えば、各分割文書データの所属カテゴリに関する情報 (たとえば、図45に示した「分割文書データを3つの 【0332】ここで、分割文書分類結果情報とは、たと

> ような種々の情報を分類結果分析の際の基礎データとし いう表の中の情報)などである。なお、利用者は上記の 数)」の項の情報)、生成された所属カテゴリ間の情報 いう表中の「代表値」および「所属データ数(分割文書 て利用することができる。 【0333】図45は、12個の分割文書データをそれ (たとえば図45に示した「分類カテゴリ間の距離」と

3つのカテゴリに分類することができる。 クトルに線形変換している) に対してたとえばクラスタ なるが、ここでは、いくつかの単語が縮退した3次元ペ 数は分類対象文書群に生起するすべての単語の種類数に 一夕の有する計量的な3次元ペクトル(ペクトルの成分 らの有する計量的特徴ベクトルを用いて3つのカテゴリ 分析手法の一つであるWard法などを適用することで に分類した場合の分類結果の生成例である。分割文書デ

特徴ベクトルの平均値(所属分割文書データの重心) に3つのカテゴリのうちのいずれか一つに属する。な お、所属カテゴリの代表値とは、所属分割文書データの 【0334】つまり、各分割文書データは図示したよう

似度に対応する)は、たとえば、図45の分割文書3に により、以下の数式から求めることができる。 ゴリ2の代表値(所属分割文書データの重心)の項の値 割文書3の値と、分割文書3の分類カテゴリであるカテ ついては、分割文書データ特徴ペクトルの項における分 【0335】また、所属カテゴリの代表値との距離(額

-2.00) ²÷ (4.00-3.66) ²) ^{1/2} = 0. [0336] ((3.00-2.66)2+(2.00

いうことになる。 のカテゴリに属する平均的分割文書との類似度が高いと 上記の所属カテゴリの代表値との距離が小さいほど、そ

散、各カテゴリにおける類似度のレンジなどさまざまな 45に示した以外にも、カテゴリ内分散やカテゴリ間分 統計量を生成することかできる。 【0337】なお、分割文書分類結果情報としては、図

記文書-分割文書対応マップと上記分割文書分類結果作 所属する割合)、分割文書データの所属文書における相 文書)、文書占有率(分割文書データの当該カテゴリに 示したように、各分類カテゴリごとに、所属する分割タ 報とを用いて、たとえば図46に示すような、上記文書 タの類似度の順位などを生成している。 対位置(順序)、所属カテゴリ内での当該分割文書デー 醚)、分割文書データの属する分割前文書データ(所順 データの分類結果情報を生成する。図46の例では、図 書データ、その類似度(所属カテゴリの代表値との距 【0338】続いて、文書分類結果生成部5006が上

文 分類結果情報から得ている。文 分類結果生成部5 割文書対応マップから、それ以外の分類結果情報は分割 【0339】なお、上記において、所属文書は文書一分

容などを分類結果情報として利用することもできる。 どさまざまな統計量、文書データや分割文 データの内 の分散、分割文書データの所属カテゴリ内での偏差値な 006は図46に示した情報以外にも、各カテゴリ内で 【0340】また、上記においては、すべての結果を分

いようにすることも可能である。 なく、グラフィカルな表現にして、利用者が理解しやす きる。また、分類結果情報をテキスト表現にするだけで 類カテゴリや文書データを単位として表現することもで 割文書データを単位とした表形式で表現しているが、分

の読みたい部分を効率的に読むことができる。 をよく理解できる。また、分割前文 (所属文書)中の ることがなく、したがって、利用者がその分類カテゴリ 意図するカテゴリとは異なるカテゴリに分類されたりす 数の話題や意味が含まれている場合に、ある特定の話題 分類結果が利用者に示されるので、一つの文 の中に複 紀分割文書との対応が利用者に示され、上記分割文書の 文書が分割され、分割文書が分類され、分割前文 と上 分割文書の位置なども示されるので、利用者は文 群中 や意味に限定されたカテゴリに分類されたり、利用者の 【0341】こうして、本実施の形態によれば、一つの

008、文書一分割文 対応マップ生成部5003によ する文書保存部(文書保存手段)5007、分割文書デ 示したように、実施の形態8の文 分類装置は、図42 形態8に係る文書分類装置の構成プロック図である。図 により模成される。 部はたとえば共有のハードディスクや半導体メモリなど り生成された文書一分割文書対応マップを保存する文書 に示した実施の形態7の構成に加え、文 データを保存 プ保存手段)5009を備えている。なお、上記各保存 一夕を保存する分割文 保存部(分割文書保存手段)5 −分割文書対応マップ保存部(文 −分割文 対応マッ 【0342】〔実施の形態8〕図47は本発明の実施の

与される場合にはこの識別子も適切な形式で保存するこ 容とともにその要素からなる計量的な特徴ベクトルを持 報を適切な形式で保存する。また、文 データが文 内 保存部5007は、文 データの内容や、文 の作成 て、個々の文書データにそれらを一意に表す識別子が付 つ場合にはこれらも保存する。文 入力部5001に 者、作成日、最終修正日などの文 データに付随する作 【0343】上記した構成により、本実施の形態の文

はこの識別子も適切な形式で保存することができる。 適切な形式で保存するとともに、計量的な特徴ベクトル 割部5002により生成される分割文 データの内容を データにそれらを一意に表す識別子が付与される場合に を持つ場合にはこれらも保存する。個々の上記分割文書 【0344】また、分割文 保存部5008は、文 分

009は、文 一分割文 対応マップ生成部5003に 【0345】また、文 一分割文 対応マップ保存部5

特開2000-285140

(36)

より生成される文 一分割文書対応マップを適切な形式で保存する。

[0346] このように、実施の形態8によれば、文書データ、分割文 データ、および文書一分割文書対応マップが保存されるので、分割文書データおよび文書一分割文書が成立 対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸股定などパラメータの異なる分類結果を効率的に求めることができる。また、文 データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者は、分類作業に対して時間的な自由度を持つことができ、過去に行った文 分類の再分析を任意の時間におこなうこともできる。

【0347】【実施の形態9】図48は本条明の実施の形態9を示す文 分類装置の構成プロック図である。図48に示したように、本実施の形態の文書分類装置は、図47に示した実施の形態のの構成に加え、分割文書分類結果生成即5005により生成された分割文書分類結果集存事段)5010を備えている。なお、上記分類結果保存事段)5010を備えている。なお、上記分類結果保存等的5010は、たとえば、共有のハードディスクや半導体メモリなどにより構成される。

【0348】このように、第3の実施の形態によれば、文 データ、分割文 データ、文書・分割文書対応マップ、および、分割文書分類結果情報が保存されるので、実施の形態8の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することかできる。また、分割文書分類結果情報が保存されることにより、分類の実行作集および分類結果の分析作業において、利用者は、時間的な自由度を持つことができ、過去に行った文書分類結果の分析をさまざまな表現形式で任意の時間におこなうこともできる。

【0349】(実施の形態10】この発明の実施の形態10では、前記各実施の形態の文書分類装置、文書分類方法において、図49に示すように、文書分割的5002により生成される複数の分割文書データ中に分割前の文 データである文書1を含む。これにより、本実施の技態では、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文 データ自体を分類した結果として得られるマクロな分類構造の融合した分類構造を得ることができる。

【0350】〔実施の形態11〕この発明の実施の形態11では、前記各実施の形態の文書分類装置、文書分類方法において、文 分割部5002は、文書データの構造情報を基に文 データを分割する。図50に、分類対象文 データかHTML形式で記述された文書の例を示す。分割をおこなう前に、図50に示したようなHTML形式の文 データから構造情報を抽出し、それらの構し形式の文 データから構造情報を抽出し、それらの構

造を用いて文書の適切な分割規則を設定することにより文書データから分割文書データを生成する。

(0351)つまり、この例では、文書データ中のタグ くし1>に着目し、「タグくし1>を持つテキストを一つの分割文書データとする」という文書を分割文書データを生成する規則とする。この規則を文書データに適用することにより図50に示したような7つの分割文書かせまされる。

【0352】上記のように、文書が、HTML、XML、SGMLなど特定の構造化文書の形式を有していない場合でも、文字の大きさ、文字の装飾、文字の色、およびフォントなどに関する情報から分割規則を生成し、分割文書を生成することもできる。また、文書データがイメージであってOCR装置などにより入力される場合には、元のイメージのレイアウト情報などを利用することにより分割規則を生成し、分割文書を生成することもにより分割規則を生成し、分割文書を生成することもでまる。

【0353】なお、文書データのすべてをいずれかの分割文書データにする必要はない。たとえば、図50に示した例では、文字列「ニューストピック(98/09/ 25)」は分割文書には採用しない。

[0354] このように、実施の形態11では、文書データから構造情報を抽出し、文書割をおこなう前に構造情報を抽出し、文書割をおこなう前に構造情報を用いて文書の適切な分割規則を設定することにより、異なった話題の分割などを適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

[0355][実施の形態12]この発明の実施の形態12では、前記実施の形態7~10の文書分類技量、文書分類方法において、図51に示すように、文書データに含まれる単語など要素を抽出する文書要素解析的(文書要素抽出手段)5011によび含量要素付随情報抽出手段)5012を購入(図51は図48に示した実施の形態9に文書要素付随情報抽出部5011、要素付随情報抽出手段)5012を加えた例で示している)、文書分割的5002が、上記文書要素解析的5011により抽出された要素、または上記要素と上記要素付随情報抽出部5011により抽出された要素、または上記要素と上記要素付随情報も出部5012により抽出された要素、または上記要素と上記要素付随情報とを用いて上記文書データを分割する。

【0356】図52に示すように、文書分割をおこなう 前に、自然書語処理手段である文書要素解析部5011 が文書データから単語などそれらの要素を抽出し、要素 付随情報抽出部5012が品詞など要素付随情報を抽出 して文書の適切な分割規則を設定するのである。なお、 上記文書要素解析部5011および要素付随情報抽出部 5012は新たに設けるのではなく、分割文書分類部5004内の同様の手段を用いることが可能である。

【0357】この実施の形態では、たとえば、図52に

示したように、文書データが特定の構造情報を持たない複数のニューストピックの集まりであり、各トピックが、単語「トピック」・「数字」・「改行コード」という文字列の後に記述されている場合で説明すると、上記のような構造が文書更素解析部5011および要素付価情報抽出部5012の抽出結果から認識され、文章の終婚を考慮して、「トピック+数字+改行コードという文字列を先頭とし、上記文字列または文書終婚記号を終婚として囲まれる文字列を一つの分割文書データとする」という分割規則が生成されることになる。

【0358】さらに詳しく説明すると、抽出された単語とその品詞情報などから、まず、名詞と改行コードのみを抽出し、つぎに、文字列「トピック+数字+改行コード」および文書終機記号を検出し、文書内でのそれらの位置を記憶する。そして、文書データに対して前記分割、規則を適用し、図52に示したような分割文書データを生成する。

たとえば、図54に示す文 データをほぼ200文字を

たは文字数と文数の両方を基に文(データを分割する。

[0359]なお、文書データのすべてをいずれかの分割文書データにする必要はなく、たとえば、図52に示した例では、文字列「ニューストピック(98/09/25)」は分割文書には採用しない。また、上記の例では、文書データから要素およびその付随情報を抽出して、大書データが場合で説明したが、要素のみを抽出してその要素情報から分割規則を設定することも可能でもよ

なうこともできるし、文字数と文数の両方を基にした文

【0360】こうして、実施の形態12によれば、文書データからそれらの要素情報などを抽出し、抽出した要素情報などを相比で文書の分割規則を設定することにより、実施の形態11と同様に、文書データの詳細な分類り、実施の形態11と同様に、文書データの詳細な分類り、実施の形態13)にの免別の実施の形態13では、前配実施の形態13)この免別の実施の形態13では、前配実施の形態7~10の文書分類装置、文書分類方法において、利用者により指示された指定範囲を分類方法において、利用者により指示された指定範囲にしたがつて文書分割部5002が文書データに対して利用者がそれぞれの分割文書の範囲を指定すると、指定にしたがつて文書分割部5002が文書分割をおこなう。

【0362】本実施の形態では、文書分割時、文書分割 翻5002がまず、画面上に、その初期状態として左右 の指示ポイントおよび領域指定ラインからなる領域指定 オブジェクトを文書の最上部に表示する。この状態で、 利用者は、マウスなどポインティングデバイスを用い て、左右どちらかの指示ポイントをドラッグし、それを 上下に移動させることにより、それぞれの分割文書の領域を選択することができる。

[0363]また、この指定時、文書分割節5002は、領域選択処理をおこなっていることを示すため、指示ポインタを無色から白色に、領域指定ラインを実験から破壊に変化させる。選択領域を決定するには、所望の位置で指示ポイントのドラッグを止めればよい。

[0364] つぎに、利用者は選択した領域を分割文とするかしないか決定する。分割領域としない場合には、それを明示的に表示するために、文 分割師5002は選択領域を図示のように頼掛け表示にさせる。

構造がわかり、かつ利用者の意図に合った文 分類をおこなうことができる。 【0366】〔実施の形態14〕この免明の実施の形態14では、前記実施の形態7~10の文 分類装置、文書分類方法において、文 データ中の文字数、文数、ま

は文書データからそれぞれの分割文 データを所望通り に選択することができるので、文 データの詳細な分類

単位として分割をおこなう。
【0367】ここで、ほぼ200文字を単位とするの【0367】ここで、ほぼ200文字を単位とするのは、正確な200文字単位としてもその終婚が句点である保証がないことから、200文字目の前後のもっとも近い句点をそれぞれの分割文 の終婚とするからである。こうして、図54に示したような分割文書が生成される。同様に、所定の文数を単位とした文 分割をおこれる。同様に、所定の文数を単位とした文 分割をおこ

事分割をおこなうこともできる。
【0368】このように、実施の形態14によれば、文で368】このように、実施の形態14によれば、文字数、文数、または文字数と文数の両方を基に文 データを分割することにより、話題の異なった内容などが異々った分割文書として分割され、分類される可能性が高くなるので、文書データの詳細な分類構造がわかる文分類をおこなうことができる。

【0369】〔実施の形態15〕この免明の実施の形態15では、前記各実施の形態の文 分類装置、文 分類方法において、文書分類結果生成部5006が分類結果情報として、文書データを示す情報および上記文 データに付随する代表的情報のみを提示する。

【0370】たとえば図55に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文 データを含んでいる文書データの、たとえば、文 データ名(文書名)を表示する。また、各文 データ名の左頭には文書アイコンを表示させ、この文 アイコンが指示されたとき、文書データの内容を表示させる。

【0371】また、各文 データ名の配置順は、カテゴリ代表値との類収度が高い分割文 データの文 データ名を先(左側)にする。また、同じ文 データから生成された複数の分割文 データが同一の分類カテゴリに属している場合には、類収度のもっとも高い分割文 データに対応する文書データ名のみを表示する。なお、上記キーワードとは出現頻度の多い単語である。

【0372】このように、実施の形態15によれば、文

(32)

分類結果が文 データを示す情報と文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要を容易に把握することができる。

【0373】(実施の形態16〕この発明の実施の形態16では、実施の形態15の文書分類結果提示に加えて、分割文 データを示す情報および上記分割文書データに付随する情報を提示する。

【0374】たとえば、図56に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文 データのたとえば文書データ名(文書名)を表示する。

【0375】また、各文 データ名の左側には文書アイコンを表示させ、この文 アイコンが指示されたとき、文 データの内容を表示させる。また、文書データ名の右側には分割文 アイコンを表示させる。なお、分割文 アイコン中には当該文書データにおける分割文書データの位置と当該文 データ中の分割文書教を表示させる。さらに、上記分割文 アイコンを指示することで文データ中の当該分割文 データを表示させることがで

【0376】また、各文 データ名の配置原はカテゴリ代表値との類似度が高い分割文書データの文書データ名を先にする。また、同じ文書データから生成された複数の分割文 データが同一の分類カテゴリに属している場合には類似度の順位がわかるようにその順位を表示させる。

【0377】このように、実施の形態16によれば、文分類結果が文書データを示す情報と文書データに付随する代表的情報、および分割文書データを示す情報と分割文 データに付随する代表的情報のみが表示されるので、利用者は文 データの詳細な分類構造の概要とともにどの分割文 が起因して当該カテゴリに分類されたかというようなことも容易にわかる。

【0378】以上、本発明の文書分類装置および文書分類方法を説明したが、この文書分類方法を実現するプログラムを着脱可能であるとともにコンピュータ読み取りグラムを着脱可能であるとともにコンピュータ読み取り可能な記録媒体に記録し、上記記録媒体を移した先の情報処理装置内で本発明によった文書分類をおこなうこともできる。

[0379]

【発明の効果】以上説明したように、請求項1の発明によれば、入力された文 データを記憶する文書記憶手段と、前記文 記憶手段により記憶された文書データの全部または一部を退択する選択手段と、前記選択手段により選択された文 データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、前記特徴抽出手段により抽出された文字列の特徴に関するデー

タに基づいて前記文書データの全部または一部を加工処理手段と、前記加工処理手段により加工処理手段と、前記加工処理手段により加工処理主なた文書データの全部または一部を出力する出力手段とを備えるため、文書の意味に保わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0380】また、請求項2の発明によれば、削配出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記文書データの全部または一部の大容に基づいて複数の項目値を設定された項目値設定手段と、前記文書データの全部さたは一部を実計する集計手段と、を備え、前元文書データの全部または一部を、項目値を少なくとも一つの軸とする要部または一部を、項目値を少なくとも一つの軸とする要形式に展開して出力するため、簡易な操作で加工処理の結果をクロス要として要すことができ、信頼の内容のは概念を第におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0381】また、請求項3の発明によれば、前配出力手段が、さらに、前配加工処理手段により加工処理された文書データの全部または一部を、前配加工処理手段により加工処理された文書データの全部または一部とより加工処理される前の文書データの全部または一部とともに出力するため、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなり、加工処理の対象範囲の決定を正確かつ容易におこなり、加工処理の対象範囲の決定を正確かつ容易におこなり、したができることから、文書の意味に張わるような分うことができることから、文書の意味に張わるような分が作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を表する。

【0382】また、請求項4の発明によれば、前記文書記憶手段が、さらに、前記加工処理手段により加工処理 された文書データの全部または一部を記憶するため、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0383】また、請求項5の発明によれば、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択するため、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業になった、単にその結果のみを出力するのではなく、情報分析作業全段にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0384】また、請求項6の発明によれば、前記文

タを記憶するため、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に保づけて把握することができることから、文書の意味に保づけるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全級にわたる支援をおになうことが可能な文書処理装置が得られるという効果をすった。

記憶手段が、さらに、前記加工処理の内容に関するデー

かを漸次的に収集することが可能な文書分類装置が得ら れにより、任意の文書集合にどのような内容が含まれる くクラスタの構造化・体系化をおこなうことができ、こ するため、クラスタを得ることができるとともに、クラ 出されたクラスタ特徴を分類体系の構成要素として記憶 類体系記憶手段が、前記クラスタ特徴算出手段により算 文書の部分集合の特徴であるクラスタ特徴を算出し、分 ラスタ特徴算出手段が、前記分類手段により生成された に基づいて文書を分類し、文書の部分集合を生成し、ク 生成手段により生成された文書特徴ベクトル間の類似度 文書特徴ベクトルを生成し、分類手段が、前記ベクトル られた言語解析情報に基づいて前記文書データに対する を得、ベクトル生成手段が、前記書語解析手段により得 段により入力された文書データを解析して言語解析情報 が、文書データを入力し、言語解析手段が、前記入力手 れるという効果を奏する。 スタ重心間の類似度等を用いて、クラスタの内容に基づ 【0385】また、請求項7の発明によれば、入力手段

スタの構造化・体系化をおこなうことができ、これによ の部分集合を選択し、分類体系記憶手段が、前記クラス 分類手段により生成された文書の部分集合の中から所望 クラスタ特徴を表示し、クラスタ選択指示手段が、前記 **示手段が、前記クラスタ特徴算出手段により算出された** 文書の部分集合の特徴であるクラスタ特徴を算出し、表 に基づいて文書を分類し、文書の部分集合を生成し、ク 生成手段により生成された文書特徴ペクトル間の類似度 文書特徴ベクトルを生成し、分類手段が、前記ベクトル られた書語解析情報に基づいて前記文書データに対する 段により入力された文書データを解析して言語解析情報 が、文書データを入力し、言語解析手段が、前記入力手 次的に収集することが可能な文書分類装置が得られると り、任意の文書集合にどのような内容が含まれるかを タのみを用いて、より操作者の意図したものに近いクラ 体系の構成要素として記憶するため、選択されたクラス 夕選択指示手段により選択された文書の部分集合を分類 ラスタ特徴算出手段が、前記分類手段により生成された を得、ベクトル生成手段が、前記言語解析手段により得 【0386】また、請求項8の発明によれば、入力手段

【0387】また、請求項9の発明によれば、請求項8の発明において、文 特徴ベクトル記憶手段が、前記べ

クトル生成手段により生成された文_特徴ペクトルを配慮し、ベクトル棒正手段が、前記文 特徴ペクトル記憶 億し、ベクトル棒正手段が、前記文 特徴ペクトル記憶 手段により記憶された女皇特徴ペクトルを、前記クラス 夕選択指示手段により選択された部分集合に属する文の文章特徴ペクトルを除去したのごりとなるように存正し、前記分別手段が、前記ペクトル棒正手段により棒正された童特徴ペクトルに基づいて文 を分類するため、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文章集合にどのような内容が含まれるかを漸次的に収集することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類装置が得られるという効果を奏することが可能な文章分類を記された。

することが可能な文書分類装置が得られるという効果を 文書集合にどのような内容が含まれるかを漸次的に収集 なクラスタを生成することができ、これにより、任意の 分類実行時に排除することができ、排除した状態で新た 結果、操作者に選択されたクラスタの形成特徴を次回の 似度に基づいて文書を分類するため、前回の分類実行の トル生成手段により生成された文 特徴ペクトル間の類 手段により修正された文 表現空間を用いて、前記べつ **づいて修正し、前記分類手段が、前記文 表現空間修正** 手段により選択された部分集合から算出する特徴量に基 度を判断する際の文書表現空間を前記クラスタ選択指示 ル記憶手段により記憶された文 特徴ベクトル間の類似 記憶し、文書表現空間修正手段が、前記文 特徴ベクト ベクトル生成手段により生成された文 特徴ベクトルを 8の発明において、文 特徴ベクトル記憶手段が、前記 【0388】また、請求項10の発明によれば、請求項

容が含まれるかを漸次的に収集することが可能な文 分 とができ、これにより、任意の文 集合にどのような内 とができ、排除した状態で新たなクラスタを生成するこ された文書表現空間を用いて、前記ペクトル修正手段に 手段により選択されたクラスタ特徴に基づいて修正し、 度を判断する際の文書表現空間を前記クラスタ選択指示 ル記憶手段により記憶された文 特徴ペクトル間の類似 記憶し、文書表現空間修正手段が、前記文 特徴ベクト 9の発明において、文書特徴ベクトル記憶手段が、前記 類装置が得られるという効果を奏する。 たクラスタの形成特徴を次回の分類実行時に排除するこ 除し、かつ、前回の分類実行の結果、操作者に選択され 文書を分類するため、既知になったクラスタの影響を排 より修正された文書特徴ベクトル間の擬弦度に基心いて 前記分類手段が、前記文書表現空間修正手段により修正 ベクトル生成手段により生成された文書特徴ベクトルを 【0389】また、請求項11の発明によれば、請求リ

【0390】また、請求項12の発明によれば、請求項8または10の発明において、選択情報付与手段が、前記分類手段により生成された文 の部分集合に所属する文 のすべてあるいは一部が選択された場合に選択され

(34)

たことを示す選択情報を付与し、前記表示手段が、前記クラスタ特徴を表示するとともに、選択情報付与手段により付与された選択情報を表示するため、多重に利用される文 の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0391】また、請求項13の発明によれば、請求項8~12の発明において、前配分類体系配億手段が、前記規則指示手段により選択された文書の部分集合に属する全部あるいは一部の文のほか、クラスタ特徴および/または操作者が作成した任意の情報を分類体系の構成要素として配億するため、クラスタの内容把超を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができ、これにより、任意の文集合にどのような内容が含まれるかを満次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0392】また、請求項14の発明によれば、文書の内容にしたがって文、群を分類する文書分類装置において、文 データ群を入力する文書入力手段と、入力されて、文 データ群の各文書に対して所定の基準に基づきさた文 データ群の各文書に対して所定の基準に基づきでの分割をおこない、一つの文書データがまで、 たれば

複数の分割文 データを生成する文書分割手段と、前記 分類されたり、利用者の意図するカテゴリとは異なるカ 果情報を生成する文 分類結果生成手段と、を備えるた づいて分割文 分類結果情報を生成する分割文書分類結 ブ生成手段と、前記分割文書データを分類する分割文書 分割文 対応マップを生成する文書一分割文書対応マッ 文 データと前記分割文 データとの対応を示す文書ー むことが可能な文 分類装置が得られるという効果を奏 るので、利用者は文 群中の読みたい部分を効率的に読 用者がその分類カテゴリをよく理解が可能で、また、分 テゴリに分類されたりすることがなく、したがって、利 場合に、ある特定の話題や意味に限定されたカテゴリに め、一つの文 の中に複数の話題や意味が含まれている 割文 分類結果情報とを用いて前記文書データの分類結 果生成手段と、前記文 一分割文書対応マップと前記分 分類手段と、前記分割文 分類手段による分類結果に基 (所属文)中の分割文書の位置なども示され

【0393】また、請求項15の発明によれば、請求項14の発明において、前記文書データを保存する文書保存手段と、前記分割文 データを保存する分割文書保存手段と、前記文 一分割文書対応マップ生成手段により生成された文 一分割文 対応マップを保存する文書・分割文 対応マップ保存手段と、を購えるため、分割文データおよび文 一分割文 対応マップを再生成することなしに、同一の文 データに対して、分類数、分類

手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることが可能で、また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者が分類作業に対して時間的な自由度を持つことが可能で、過去に行った文書分類の再分析を任意の時間間におこなうことも可能な文書分類終度が得られるという効果を奏する。

[0394]また、請求項16の発明によれば、請求項15の発明におれば、請求項15の発明におれて、前記分割文書分類結果生成手段により生成された分割文書分類結果性報を保存する分割文書分類結果保存手段を備えるため、請求項15の発明の効果に加え、一度分類を実行すれば、その分類結果をデキスト表現や美表現やグラフ表現などさまざまな形式で表現することが可能で、また、分割文書分類結果情報が表現することが可能で、また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果保存とが可能で、過去に行った文書分類結果の再分析をきまとが可能で、過去に行った文書分類結果の再分析をきまとまり表現形式で任意の時間におこなうことも可能な文書分類結果が得られるという効果を事する。

【0395】また、講求項17の発明によれば、請求項17の発明によれば、請求項14~16の発明において、前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むため、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることが可能な文書分類装置が得られるという効果を奏する。

【0396】また、講求項18の発明によれば、請求項14~17の発明において、前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたため、異なった話題の分割等を適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことが可能な文書分類接進がわかる文書分類を適切におこなうことが可能な文書分類接進が待られるという効果を奏する。

[0397]また、請求項19の発明によれば、請求項14~17の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記文書等素抽出手段と、り抽出された要素付配情報抽出手段と、を備え、前記文書分割手段が、前記文書要素抽出手段により抽出された要素付配情報抽出手段により抽出された要素が記憶情報を抽出手段により抽出された要素が記憶情報と各用いて前記文書データを分割する構成にしたため、文書データの詳細な分類構造がわかる文章分類を追切におこなうことが可能な文書分類装置が得られるという効果を奏する。

【0398】また、請求項2のの発明によれば、請求項 14~17の発明において、前記文書分割手段が、指示 された指定範囲にしたがって文 データの分割をおこな う構成にしたため、利用者の意図に合い、かつ文 デー

タの詳細な分類構造がわかる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。 [0399]また、請求項21の免明によれば、請求項14~17において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたため、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類 構造がわかる文書分類をおこなうことが可能な文書分類 構造がわかる文書分類をおこなうことが可能な文書分類 接置が得られるという効果を奏する。

【0400】また、請求項22の免明によれば、請求項14~21の免明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付配が、文書データを示す情報および前記文書データに付配する代表的情報を、分類結果情報として抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することが可能な文書分類装置が得られるという効果を表する。

【0401】また、請求項23の発明によれば、請求項22の発明におれば、請求項22の発明において、前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造とともにどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる文書分類装置が得られるという効果を奏する。

[0402]また、請求項24の発明によれば、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程と、前記文書記憶工程と、前記文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を透析工程により選択された文書データの全部または一部を加工処理する加工処理を持た大学判の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理された文書データの全部または一部を加工処理された文書データの全部または一部を加工処理された文書データの全部または一部を加工処理された文書データの全部または一部を加工処理された文書データの全部または一部を加する出力工程と、を含むので、文書の意味に保わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0403】また、請求項25の発明によれば、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力するので、簡易な操作で加工処理を形式に展開して出力するので、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の的規定を容易におこなうことができることから、文の意

味に保わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援 を出力するのではなく、情報分析作業全般にわたる支援 をおこなうことが可能な文 処理方法が得られるという か思え奏する

【0404】また、講求項26の発明によれば、前記出力工程が、さらに、前記加工処理工程により加工処理工程により加工処理工程により加工処理大会が大学データの全部または一部を、前記加工処理大会に出力するので、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることかで、全ることから表している。大の意味に係わるような分析で美において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことがなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を要する。

【0405】また、請求項27の免明によれば、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶するので、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその核果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文 処理方法が得られたる支援をおこなうことが可能な文 処理方法が得られるという効果を奏する。

【0406】また、請求項28の免明によれば、前記選択工程が、さらに、前記出力工程により出力された文データの全部または一部を選択するので、出力工程により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることがら、文の意味に係わるような分析作業ができることができ、変にないまなく、情において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0407】また、請求項29の毎明によれば、前配文章記憶工程が、さらに、前配加工処理の内容に関するデータを記憶するので、加工処理の内容に関するデータの例表を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を問題づけて把煙することができることから、文書の意味に係かるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果をなった。

【0408】また、請求項30の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文 データを解析して言語解析撰により報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文 データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文 特徴ベクトル間の類似度に基づいて文 を分類し、文 の部分集合を生成し、

いて分類体系の構成要素を生成するので、選択されたク 望の部分集合を選択し、分類体系生成工程が、前記クラ 記分類工程により生成された文書の部分集合の中から所 たクラスタ特徴を表示し、クラスタ選択指示工程が、前 表示工程が、前記クラスタ特徴算出工程により算出され た文 の部分集合の特徴であるクラスタ特徴を算出し、 度に基づいて文 を分類し、文書の部分集合を生成し、 ル生成工程により生成された文書特徴ベクトル間の類似 る文 特徴ベクトルを生成し、分類工程が、前記ベクト 得られた言語解析情報に基づいて前記文書データに対す 程が、文書データを入力し、宮語解析工程が、前記入力 を漸次的に収集することが可能な文書分類方法が得られ により、任意の文書集合にどのような内容が含まれるか クラスタの構造化・体系化をおこなうことができ、これ **ラスタのみを用いて、より操作者の意図したものに近い** スタ選択指示工程により選択されたクラスタ特徴に基づ クラスタ特徴算出工程が、前記分類工程により生成され 報を得、ベクトル生成工程が、前記書語解析工程により 工程により入力された文 データを解析して言語解析側 【0409】また、請求項31の発明によれば、入力エ

合にどのような内容が含まれるかを漸次的に収集するこ スタを生成することができ、これにより、任意の文書集 で、既知になったクラスタの影響を排除した新たなクラ 夕選択指示手段により選択された部分集合に属する文書 3 1の発明において、ベクトル修正工程が、前記クラス とが可能な文書分類方法が得られるという効果を奏す された文書特徴ベクトルに基づいて文書を分類するの し、前記分類工程が、前記ベクトル修正工程により修正 の文 特徴ベクトルを除去したのこりとなるように修正 【0410】また、請求項32の発明によれば、請求項

修正工程により修正された文 表現空間を用いて、前記 に基づいて修正し、前記分類工程が、前記文書表現空間 指示工程により選択された部分集合から算出する特徴量 類似度を判断する際の文 表現空間を前記クラスタ選択 クトル生成工程により生成された文書特徴ベクトル間の 31の発明において、文 表現空間修正工程が、前記ペ ベクトル生成手段工程により生成された文 特徴ベクト 【0411】また、請求項33の発明によれば、請求項

> 類実行の結果、操作者に選択されたクラスタの形成特徴 次的に収集することが可能な文書分類方法が得られると り、任意の文書集合にどのような内容が含まれるかを漸 態で新たなクラスタを生成することができ、これによ を次回の分類実行時に排除することができ、排除した状 ル間の類似度に基づいて文書を分類するので、前回の分

クトル生成工程により生成された文書特徴ベクトル間の の類似度に基づいて文書を分類するので、既知になった 指示工程により選択された部分集合から算出する特徴量 類似度を判断する際の文書表現空間を前記クラスタ選択 32の発明において、文書表現空間修正工程が、前記べ ることが可能な文書分類方法が得られるという効果を奏 クラスタを生成することができ、これにより、任意の文 類実行時に排除することができ、排除した状態で新たな 果、操作者に選択されたクラスタの形成特徴を次回の分 クラスタの影響を排除し、かつ、前回の分類実行の結 ベクトル修正工程により修正された文書特徴ベクトル間 修正工程により修正された文書表現空間を用いて、前記 に基づいて修正し、前記分類工程が、前記文書表現空間 **書集合にどのような内容が含まれるかを漸次的に収集す** 【0412】また、請求項34の発明によれば、請求項

記クラスタ特徴を表示するとともに、選択情報付与工程 により付与された選択情報を表示するので、多重に利用 れたことを示す選択情報を付与し、前記表示工程が、前 前記分類工程により生成された文書の部分集合に所属す 31または33の発明において、選択情報付与工程が、 ことが可能な文書分類方法が得られるという効果を奏す 集合にどのような内容が含まれるかを漸次的に収集する 別性を向上させることができ、これにより、任意の文書 される文書の識別性および一度も選択されない文書の識 る文書のすべてあるいは一部が選択された場合に選択さ 【0413】また、請求項35の発明によれば、請求項

を漸次的に収集することが可能な文書分類方法が得られ で、分類体系の利用価値を向上させることができ、これ つ、操作者独自の分類体系を簡易に生成できることの 集合に所属する文書群の全部あるいは一部および/また 前記選択指示工程により選択されたクラスタ特徴のほ るという効果を奏する。 により、任意の文書集合にどのような内容が含まれるか を生成するので、クラスタの内容把握を容易にし、か は操作者が作成した情報に基づいて分類体系の構成要素 か、前記文書の部分集合の中から選択された文書の部分 31~35の発明において、前記分類体系生成工程が、 【0414】また、請求項36の発明によれば、請求項

る特定の話題や意味に限定されたカテゴリに分類された 【0415】また、請求項37の発明によれば、一つの の中に複数の話題や意味が含まれている場合に、あ

×

な文書分類方法が得られるという効果を奏する。 者は文書群中の読みたい部分を効率的に読むことが可能 属文書)中の分割文書の位置なども示されるので、利用 分類カテゴリをよく理解できる。また、分割前文書(所 類されたりすることがなく、したがって、利用者がその り、利用者の意図するカテゴリとは異なるカテゴリに分

ログラムを機械読み取り可能となり、これによって、請 24~37のいずれか一つに記載された方法をコンピュ ことが可能な記録媒体が得られるという効果を奏する。 水項24~37の動作をコンピュータによって実現する 一夕に実行させるプログラムを記録したことで、そのブ

【0416】また、請求項38の免明によれば、請求項

構成する情報処理システム全体のハードウエア構成を示 す説明図である。 【図1】この発明の実施の形態1による文書処理装置を

エア構成を示す説明図である。 構成する情報処理システムにおけるサーバーのハードウ 【図2】この発明の実施の形態 1 による文書処理装置を

構成する情報処理システムにおけるクライアントのハー ドウエア構成を示す説明図である。 【図3】この発明の実施の形態1による文書処理装置を

構成を機能的に示すブロック図である。 【図4】この発明の実施の形態1による文書処理装置の

項目名と項目値の関係を示す説明図である。 【図5】この発明の実施の形態1による文書処理装置の

文書記憶部に記憶された文書のデータ構造を示す説明図 【図6】この発明の実施の形態1による文書処理装置の

の出力部による画面表示の別の例を示す説明図である。

【図28】この発明の実施の形態1による文 処理装置 【図27】この発明の実施の形態1による文書処理装置 の出力部による画面表示の別の例を示す説明図である。

文書記憶部に記憶された文書の別のデータ構造を示す説 【図7】この発明の実施の形態1による文書処理装置の

出力部による画面表示の例を示す説明図である。 【図8】この発明の実施の形態1による文書処理装置の

出力部による画面表示の別の例を示す説明図である。 【図9】この発明の実施の形態1による文書処理装置の

の出力部による画面表示の別の例を示す説明図である。 の特徴抽出部によりおこなわれる抽出処理の内容の一覧 【図11】この発明の実施の形態1による文書処理装置 【図10】この発明の実施の形態1による文書処理装置

の加工処理部によりおこなわれる加工処理の内容の一覧 を示す説明図である。 を示す説明図である。 【図12】この発明の実施の形態1による文書処理装置

の各項目の特徴ベクトルを示す説明図である。 【図13】この発明の実施の形態1による文書処理装備

の単語とその単語IDごとの出現回数を示す説明図であ 【図14】この発明の実施の形態1による文書処理装置

【図15】この発明の実施の形態1による文 処理装置

の出力部によるクロス表作成のための指示画面を示す説 の出力部による画面表示の別の例を示す説明図である。 【図16】この発明の実施の形態1による文書処理装置

(36)

特開2000-285140

の出力部による分類処理の結果が表示された別のクロス 示す説明図である。 【図18】この発明の実施の形態1による文 処理装置

の出力部による分類処理の結果が表示されたクロス表を

【図17】この発明の実施の形態1による文 処理装置

喪を示す説明図である。 【図19】この発明の実施の形態1による文 処理装置

のクロス表の出力手順を示すフローチャートである。 の出力部の詳細な構成を示すブロック図である。 の文書記憶部の詳細な構成を示すプロック図である。 の出力部による画面表示の別の例を示す説明図である。 の出力部による画面表示の別の例を示す説明図である。 の出力部による画面表示の別の例を示す説明図である。 の出力部による画面表示の別の例を示す説明図である。 【図25】この発明の実施の形態1による文 処理装置 【図22】この発明の実施の形態1による文 処理装置 【図21】この発明の実施の形態1による文書処理装置 【図20】この発明の実施の形態1による文 処理装置 【図26】この発明の実施の形態1による文 処理装置 【図24】この発明の実施の形態1による文 処理装置 【図23】この発明の実施の形態1による文 処理装置

の構成を機能的に示すブロック図である。 の文書処理の一連の手順を示すフローチャートである。 の出力部による画面表示の別の例を示す説明図である。 のクラスタ特徴表示部の表示の一例を示す説明図であ 【図31】この発明の実施の形態2による文 分類装置 【図30】この発明の実施の形態2による文 分類装置 【図29】この発明の実施の形態11による文 処理装置

の一連の処理の手順を示すフローチャートである。 【図33】この発明の実施の形態3による文 分類装置 【図32】この発明の実施の形態2による文書分類装置

の構成を機能的に示すプロック図である。 −連の処理の手順を示すフローチャートである。 【図34】この発明の実施の形態3よる文 分類装置の 【図35】この発明の実施の形態4による文書分類装置

の構成を機能的に示すブロック図である。 【図36】この発明の実施の形態4よる文 分類装置の

の構成を機能的に示すプロック図である。 **-連の処理の手順を示すフローチャートである。** 【図37】この発明の実施の形態5による文書分類装置

【図38】この発明の実施の形態5よる文 分類装置の

の構成を機能的に示すブロック図である。 の分類結果記憶部において設けられたテーブルを示す説 一連の処理の手順を示すフローチャートである。 【図40】この発明の実施の形態6による文書分類装置 【図39】この発明の実施の形態6による文書分類装置 311 313 400 バス スキャナ マウス サーホーブ

(37)

特開2000-285140

(38)

特開2000-285140

の選択情報付与部の処理の手順を示すフローチャートで 【図41】この発明の実施の形態6による文書分類装置

の構成ブロック図である。 【図42】この発明の実施の形態7を示す文書分類装置

および文 分類方法の説明図である。 【図43】この発明の実施の形態7による文書分類装置

および文 分類方法の他の説明図である。 【図44】この発明の実施の形態7による文書分類装置

および文 分類方法の他の説明図である。 および文 分類方法の他の説明図である。 【図46】この発明の実施の形態7による文書分類装置 【図45】この発明の実施の形態7による文書分類装置

の構成ブロック図である。 の構成ブロック図である。 【図48】この発明の実施の形態9による文書分類装置

置および文書分類方法の説明図である。 【図49】この発明の実施の形態10による文書分類装

置および文 分類方法の説明図である。 【図51】この発明の実施の形態12による文書分類装 【図50】この発明の実施の形態11による文書分類装

置の構成ブロック図である。 【図52】この発明の実施の形態12による文書分類装

置および文 分類方法の説明図である。 置および文 分類方法の説明図である。 【図53】この発明の実施の形態13による文書分類装

置および文 分類方法の説明図である。 【図54】この発明の実施の形態14による文書分類装

置および文 分類方法の説明図である。 【図55】この発明の実施の形態15による文書分類装

置および文 分類方法の説明図である。 【図56】この発明の実施の形態16による文書分類装

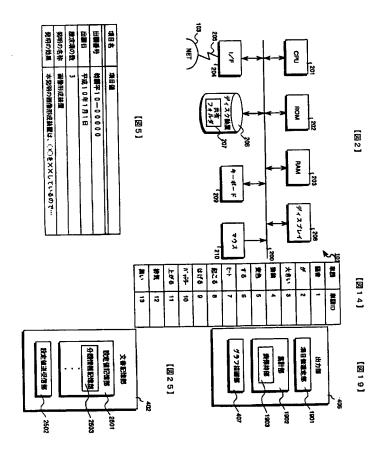
[符号の説明]

201 204 103 102 101 CPU ネットワーク クライアント 4-11-1/5

308 301 306 206 CPU ハードディスク ディスプレイ ディスク装置

【図47】この発明の実施の形態8による文書分類装置 406 405 404 401 409 408 407 403 402 3005 3002 2802 2703 2503 2502 1800 マウスポインタ 1700 1600 410 特徴ベクトル生成部 2604 2501 2401 1902 1901 1801 3301, 3501, 3701 ベクトル記憶部 3011 3010 3009 3008 3007 3006 3004 3003 3001 1903 4000 3901 3502.3703 文書表現空間修正部 3302. 3702 ベクトル修正部 5002 5001 出七卷 選技部 入力部 加工処理結果保持部 解析部 加工処理部 特徵抽出的 文書記憶部 グラフ描画部 無計割 内容表示画面 指示画面 入力部 表保持部 項目值選定部 クロス表 分類部 ベクトル生成部 分類情報表示画面 設定值送受信部 分割文 分類結果生成部 クラスタ特徴表示部 分類結果記憶部 分類パラメータ指示部 言語解析部 表示領域 問い合わせ画面 設定值記憶部 分割文 分類部 文書分割部 文書入力部 選択情報付与部 カーソル 分類体系閲覧操作部 クラスタ選択指示部 分類体系記憶部 クラスタ特徴算出部 分類情報記憶部 棒グラフ表示領域 文書-分割文書対応マップ生成部 テーブル

> 5009 文書一分割文書対応マップ保存部 5007 5006 5008 分割文書保存部 文書保存部 文書分類結果生成部 [28] 5010 分割文書分類結果保存部 5012 要素付随情報抽出部 5011 文書要素解析部 [図12] 算術包理 被崇起国 を変め **斯洛氏联局小學** 最大值排出规理 **代表语数出现型** 甘人物人名国 加工规理内容

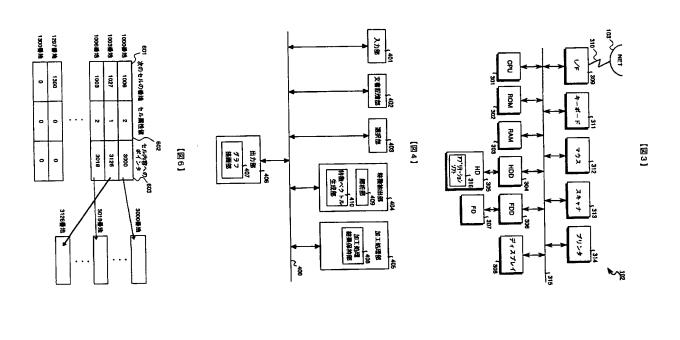




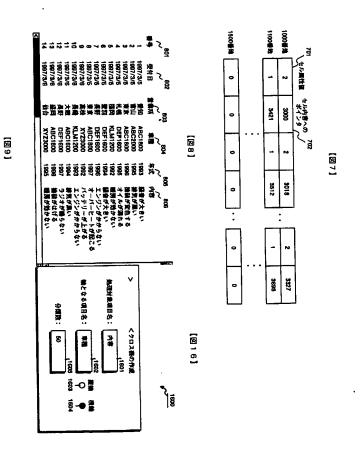
(**4**0)

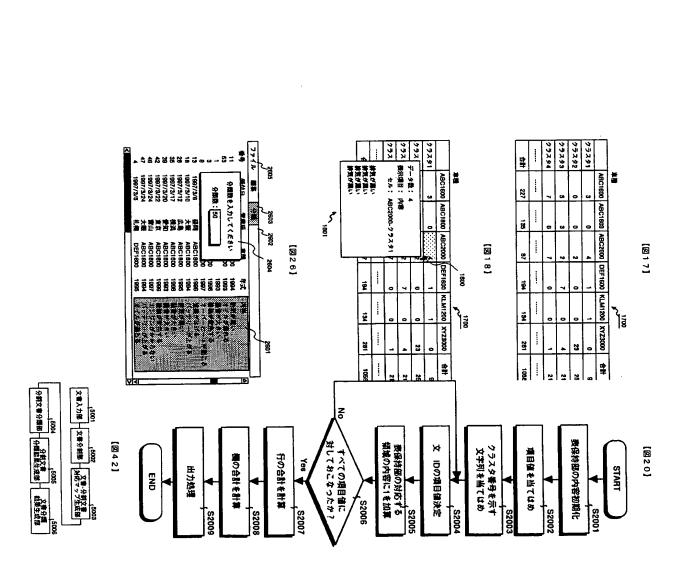
特開2000-285140

(39)



(図13)





対象とする文字列に含まれる単語間の関係の情報 対象とする文字列に含まれる単語それぞれの品詞 対象とする文字列に含まれる単層数 対象とする文字列に含まれる文の間の関係 対象とする文字列に含まれる文の文節数 対象とする文字列に含まれる文の文字数 対象とする文字列に含まれる文の数 対象とする文字列に含まれる単語それぞれの出現四数 対象とする文字列に含まれる単語の文字数 対象とする文字列に含まれる単語 抽出炮彈內容 本語 ABC1800 ABC2000 ABC1800 DEF1800 MLM1200 DEF1600 DEF1600 ABC1800 ABC1600 ABC1600 ABC1600 [図15]

ABC1800 ABC1800

[図10]

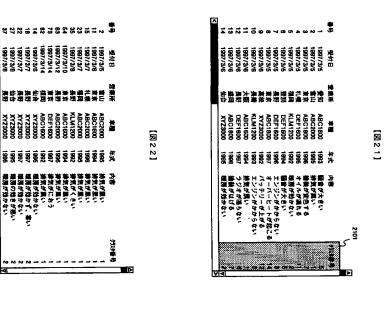
[図11]

特開2000-285140

(41)

(42)

特開2000-285140

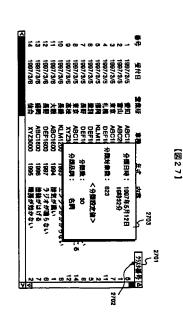


製件目 1997/3/5 1997/3/5 1997/3/5 1997/3/5 1997/3/5 1997/3/6 1997/3/6 1997/3/6 1997/3/6 <分類設定値>分類数:30 分類品詞:名詞

[图28]

受付日 1997/3/6 1997/3/6 1997/3/7 1997/3/7 1997/3/1 1997/3/14 1997/3/14 1997/3/1 1997/3/1 1997/3/1 1997/3/6 1997/3/6 1997/3/6 1997/3/6

[図23]



97 22 23 24 25 26 27 27 27 愛甘田 貨業形 BC2000 ABC1600 KLM1200 DEF1600 事務 = 推出 \$

(44)

【図24】

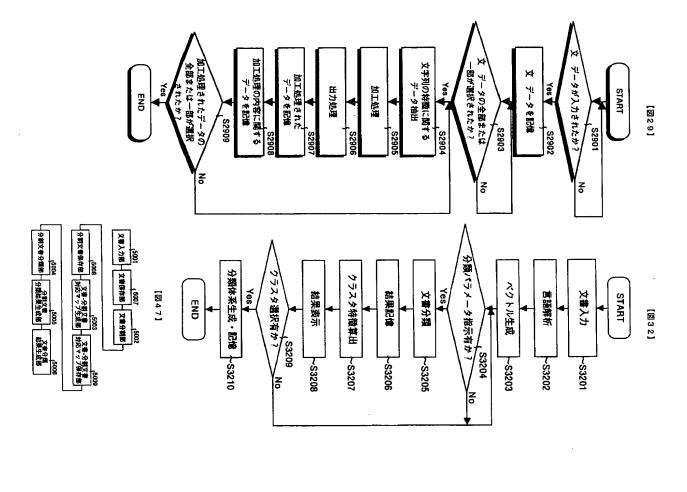
(43)

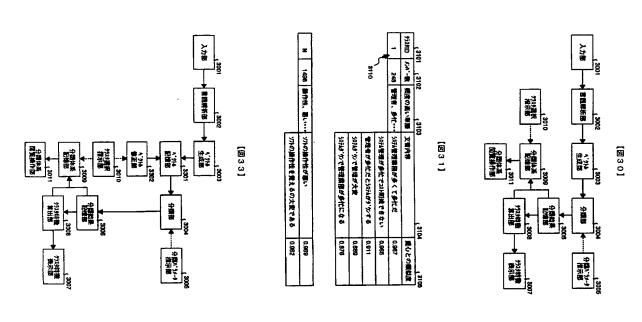
特開2000-285140

特開2000-285140

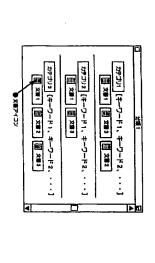


(45)





(47)



類リ近し処理するか?

END

ベクァラ奈川

分類体系生成・記憶 ∼S3411

クラスタ選択有か?

S3410

S

クラスタ特徴算出

~S3408

結果表示

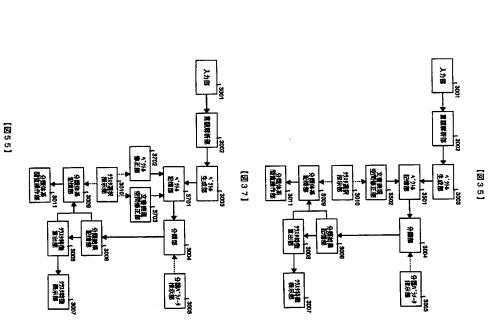
~S3409

結果記憶

~S3407

文書分類

~S3406



分類パラメータ指示

S3405

ベクトラ記録

~S3404

ベクトル生成

~S3403

宫語解析

~S3402

文書入力

~S3401

START

【図34】

(48)

(49)

